Universidade de Brasília
Faculdade de Tecnologia
Departamento de Engenharia Mecânica

Integração Multissensor de Câmera e Radar para Detecção de Objetos para desenvolvimento de Sistemas Avançados de Auxílio à Condução

Hachid Habib Cury

DISSERTAÇÃO DE MESTRADO

PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS MECATRÔNICOS

Brasília

2025

Universidade de Brasília
Faculdade de Tecnologia
Departamento de Engenharia Mecânica

Integração Multissensor de Câmera e Radar para Detecção de Objetos para desenvolvimento de Sistemas Avançados de Auxílio à Condução

Hachid Habib Cury

Dissertação de Mestrado submetida ao Departamento de Engenharia Mecânica da Universidade Brasília como parte dos requisitos necessários para a obtenção do grau de Mestre

Orientador: Prof. Dr. Evandro Leonardo Silva Teixeira

Brasília
2025

Cury, Hachid Habib.
C769i Integração Multissensor de Câmera e Radar para Detecção de Objetos para desenvolvimento de Sistemas Avançados de Auxílio à Condução / Hachid Habib Cury; orientador Evandro Leonardo Silva Teixeira. -- Brasília, 2025.
89 p.

Dissertação de Mestrado (Programa de Pós-Graduação em Sistemas Mecatrônicos) -- Universidade de Brasília, 2025.

1. Fusão sensorial. 2. Detecção de objetos. 3. Camera. 4. Radar. I. Teixeira, Evandro Leonardo Silva , orient. II. Título

Universidade de Brasília
Faculdade de Tecnologia
Departamento de Engenharia Mecânica

Integração Multissensor de Câmera e Radar para Detecção de Objetos para desenvolvimento de Sistemas Avançados de Auxílio à Condução

Hachid Habib Cury

Dissertação de Mestrado submetida ao Departamento de Engenharia Mecânica da Universidade Brasília como parte dos requisitos necessários para a obtenção do grau de Mestre

Trabalho aprovado. Brasília, 25 de agosto de 2025:

Prof. Dr.EVANDRO LEONARDO SILVA TEIXEIRA, UnB/FGA
Orientador

Prof. Dr. RENATO VILELA LOPES, UnB/FGA
Examinador interno

Prof. Dr. ABEL GUILHERMINO DA SILVA FILHO, UFPE
Examinador externo

Prof. Dr. DIVANILSON RODRIGO DE SOUSA CAMPELO
Examinador externo

Brasília
2025

Dedico este trabalho ao meu amado filho, Yohan Khaled Lins Cury.
Que mesmo antes de tê-lo em meu colo, me fazia prosseguir,
como se sua pequena existência me lembrasse do porquê não desistir,
inspirando-me, todos os dias, a ser alguém melhor para ele e por ele existir.

Agradecimentos

Agradecimentos

À minha esposa, Tayná Lins Portela Cury, pelo apoio incondicional ao longo destes anos de pesquisa, pela compreensão nos momentos de ausência e pelo cuidado dedicado a mim e, sobretudo, ao nosso filho.

Aos meus pais, Rita de Cássia Souza Cury e Rabib de Souza Cury, pela dedicação em minha formação pessoal e pelos valores que me permitiram chegar até aqui.

Ao meu orientador, Evandro Leonardo Silva Teixeira, pela confiança depositada, pela orientação competente e pela paciência demonstrada durante todas as etapas de elaboração deste trabalho.

Às minhas irmãs, Samira de Souza Cury e Madiha de Souza Cury, bem como aos meus sogros, Maria Ozana de Souza Lins e Wilson Costa Portela, pela fundamental rede de apoio e pelo cuidado com nosso filho.

Ao meu sobrinho, Sayd Moreira Cury, por todos vezes que precisou acordar de madrugada para me conceder acesso à VPN.

Por fim, à FUNDEP e ao projeto SegurAuto, pelo apoio financeiro indispensável à execução deste trabalho.

“Os olhos só enxergam o que a mente está preparada para compreender.”
(Henri Bergson)

Resumo

A percepção precisa da cena de trânsito é essencial para a segurança e a eficácia dos Sistemas Avançados de Assistência ao Condutor (ADAS), bem como para a transição rumo à condução autônoma. Este trabalho investiga a fusão de dados entre sensores de câmera e radar automotivo, utilizando a base de dados nuScenes, com foco na aplicação de técnicas de fusão nos níveis baixo e médio, tendo como referência comparativa o detector de objetos Faster R-CNN. Inicialmente, foi implementada a fusão de baixo nível por meio da *Radar Region Proposal Network* (RRPN), na qual o radar é empregado como sensor principal na geração de regiões de interesse. No entanto, essa abordagem apresentou desempenho inferior ao do detector baseado exclusivamente em câmera, uma vez que objetos não detectados pelo radar não são processados pela rede neural, comprometendo a robustez da detecção. Os resultados indicaram que mais da metade das bicicletas e motocicletas anotadas na base nuScenes não possuem qualquer ponto de radar associado; no caso dos pedestres, apenas cerca de 20% apresentam ao menos um ponto detectável por esse sensor.

Em seguida, foi avaliada a fusão de características (médio nível), com a implementação do módulo Spatial Attention Fusion (SAF) na arquitetura da rede. Os resultados demonstraram melhorias consistentes nas métricas de desempenho, com destaque para ganhos de 1.64% em AP75, 0.96% em AP50, 0.80% em AR e 1.36% em APs, indicando maior precisão na localização das caixas delimitadoras e na detecção de pequenos objetos. Esses avanços validam o potencial da fusão sensorial em nível de características como estratégia eficaz para aprimorar a percepção em sistemas autônomos.

Palavras-chave: Fusão sensorial. Detecção de objetos. Camera. Radar.

Abstract

Accurate traffic scene perception is essential for the safety and effectiveness of Advanced Driver Assistance Systems (ADAS), as well as for the transition toward autonomous driving. This work investigates data fusion between camera and automotive radar sensors using the nuScenes database, focusing on the application of low- and mid-level fusion techniques, using the Faster R-CNN object detector as a benchmark. Initially, low-level fusion was implemented using the Radar Region Proposal Network (RRPN), in which radar is used as the primary sensor in generating regions of interest. However, this approach underperformed the camera-only detector, since objects not detected by radar are not processed by the neural network, compromising detection robustness. The results indicated that more than half of the bicycles and motorcycles annotated in the nuScenes database do not have any associated radar points; in the case of pedestrians, only about 20% have at least one point detectable by this sensor.

Subsequently, mid-level feature fusion was evaluated through the implementation of the Spatial Attention Fusion (SAF) module within the network architecture. The results showed consistent improvements across performance metrics, with notable increases of 1.64% in AP75, 0.96% in AP50, 0.80% in AR, and 1.36% in APs, indicating greater accuracy in bounding box localization and enhanced detection of small objects. These advances validate the potential of feature-level sensor fusion as an effective strategy to improve perception in autonomous systems

Keywords: Sensor Fusion. Object Detection. Câmera. Radar.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ACC	Adaptive Cruise Control	19
ADAS	Sistemas Avançados de Assistência ao Condutor	7
AP	Average Precision	43
AR	Average Recall	43
BBOX	Bounding Boxes	48
BEV	Bird's Eye View	22
BSD	Blind Spot Detection	19
CNN	Convolutional Neural Network	38
COCO	Common Objects in Context	45
EBA	Emergency Brake Assist	19
FCN	Fully Connected Network	28
FMCW	Frequency Modulated Continuous Wave	22
FPN	Feature Pyramid Network	31
IOU	Intersection over Union	30
LDW	Lane Departure Warning	19
LiDAR	Light Detection and Ranging	15
RADAR	Radio Detection and Ranging	15
RCS	Radar Cross Section	39
RCTA	Rear Cross Traffic Alert	19
ROI	Region of Interest	24
RPN	Region Proposal Network	28
RRPN	Radar Region Proposal Network	7
SAF	Spatial Attention Fusion	7
ToF	Time-of-Flight	21
YOLO	You Only Look Once	29

Sumário

1 Introdução

O crescente aumento no volume de tráfego nas rodovias, aliado à ocorrência frequente de congestionamentos, à presença de sinalizações ambíguas e à constante pressão imposta pelas condições do trânsito, demanda o desenvolvimento de soluções capazes de aprimorar a segurança, a eficiência e a comodidade na condução veicular. Nesse contexto, os *Sistemas Avançados de Assistência ao Condutor* (ADAS) têm recebido atenção significativa por parte da comunidade científica internacional.

Um dos pilares para o desenvolvimento de sistemas ADAS é a detecção robusta e em tempo real de objetos presentes no ambiente rodoviário. Considerando que as condições de condução em estradas são frequentemente complexas e imprevisíveis, é necessário que os veículos estejam equipados com diferentes tipos de sensores capazes de fornecer uma percepção confiável e abrangente do entorno do veículo (YU, Z. et al., 2018). Os sensores são responsáveis pela coleta de dados que alimentam os sistemas computacionais embarcados, os quais auxiliam nas decisões relacionadas à direção, frenagem e controle de velocidade (KOCIC; JOVICIC; DRNDAREVIC, 2018). Entre os sensores mais utilizados na percepção do ambiente rodoviário estão o RADAR (*Radio Detection and Ranging*), o LiDAR (*Light Detection and Ranging*) e as câmeras.

A importância da percepção multissensorial ficou evidente em 2016, quando ocorreu, na Flórida (EUA), o primeiro acidente fatal envolvendo um veículo equipado com o sistema Autopilot da Tesla. A investigação indicou que o módulo de percepção visual interpretou erroneamente a carroceria branca de um caminhão como parte do céu claro, falhando na identificação do veículo pesado (LIU, Z. et al., 2022). Embora o sistema já integrasse múltiplos sensores, incluindo radar, sua lógica de fusão descartou os sinais potencialmente relevantes. Esse incidente evidenciou as limitações do uso isolado de sensores e os desafios da integração multissensorial, ressaltando a importância da fusão adequada de dados para maior confiabilidade na detecção de obstáculos e na compreensão do ambiente rodoviário.

A fusão de dados combina informações provenientes de diferentes sensores com o objetivo de explorar seus pontos fortes e atenuar suas limitações. Além do reconhecimento ambiental, essas tecnologias devem considerar fatores como tempo de resposta, custo e disponibilidade dos sensores para produção em larga escala, bem como a robustez em condições meteorológicas adversas. De modo geral, ao se aplicar a fusão de dados, busca-se alcançar benefícios como redundância e complementaridade de informações, melhoria na resposta temporal, tolerância a falhas e redução de custos (DARMS et al., 2010).

Dentre os principais sensores utilizados para a percepção do ambiente de trânsito, as câmeras se destacam pelo baixo custo, pela riqueza de informações e pela facilidade na classificação de objetos. No entanto, apresentam limitações, como sensibilidade a variações nas condições de iluminação e dificuldade em obter informações tridimensionais dos alvos (LIU, Z. et al., 2022). Por outro lado, os radares são capazes de detectar objetos a distâncias significativamente maiores e são altamente robustos em condições climáticas adversas. Além disso, fornecem informações precisas sobre a velocidade dos objetos detectados, permitindo prever sua trajetória e deslocamento (NABATI; HARRIS; QI, 2021). Ainda assim, os radares apresentam limitações, como a baixa densidade dos pontos de detecção, o que dificulta a estimativa de informações geométricas e a classificação precisa dos objetos (LIU, Y. et al., 2022).

A fusão sensorial entre câmera e radar oferece vantagens relevantes, pois esses sensores são complementares e amplamente utilizados na percepção automotiva. Essa integração alia a alta resolução lateral das câmeras à robustez do radar frente a variações de iluminação e condições climáticas, além de apresentar menor custo de produção em comparação aos sensores LiDAR.

1.1 Justificativa

A percepção precisa do ambiente é um dos pilares fundamentais para a operação segura e eficiente dos sistemas ADAS e veículos autônomos. Para atingir esse objetivo, os sistemas modernos de percepção embarcada recorrem à integração de múltiplos sensores, cujas características são, em grande parte, complementares. As câmeras fornecem informações visuais ricas em detalhes espaciais e semânticos, enquanto os radares oferecem medições confiáveis de distância e velocidade, mesmo em condições adversas de iluminação ou clima. No entanto, quando utilizados de forma isolada, esses sensores apresentam limitações significativas: as câmeras são sensíveis a variações de iluminação, e os radares, embora robustos, possuem baixa resolução espacial. Nesse contexto, a fusão sensorial surge como uma estratégia promissora para combinar as vantagens individuais dos sensores e mitigar suas limitações, ampliando a confiabilidade e a robustez dos sistemas de percepção.

A base de dados nuScenes foi escolhida por sua ampla adoção na comunidade científica, o que possibilita a comparação direta dos resultados com diferentes métodos de fusão para detecção 2D. Além de oferecer dados multissensoriais sincronizados e anotados em cenários reais sob diversas condições meteorológicas, mostra-se particularmente adequada a este estudo. Para a tarefa de detecção, adotou-se o Faster R-CNN, um detector clássico e amplamente consolidado na literatura, reconhecido pela alta precisão em cenários complexos. Embora existam arquiteturas mais recentes e otimizadas para execução em tempo real, a escolha do Faster R-CNN se justifica por sua estabilidade e pelo uso recorrente

como baseline em pesquisas de visão computacional. O objetivo não é comparar diferentes detectores, mas avaliar o desempenho do mesmo algoritmo apenas com imagens de câmera e, posteriormente, com fusão câmera-radar, isolando os efeitos da integração sensorial.

No que se refere à fusão entre os dados da câmera e do radar, foram selecionados os métodos *Radar Region Proposal Network* (RRPN) e *Spatial Attention Fusion* (SAF). Essa escolha fundamenta-se no levantamento do estado da arte, no qual essas abordagens apresentaram a melhor performance em seus respectivos níveis de fusão, configurando-se, portanto, como representativas tanto da fusão em nível de dados (baixo nível) quanto da fusão em nível de características (médio nível). A opção por não incluir também a fusão em nível de decisão (alto nível) deveu-se não apenas ao curto tempo disponível para desenvolver as três estratégias, mas igualmente ao fato de que, no levantamento realizado, não foram identificadas implementações desse tipo que atendessem aos critérios definidos.

Dessa forma, este trabalho se justifica pela necessidade de investigar e desenvolver técnicas eficazes de fusão de dados entre sensores de câmera e radar automotivo, com foco especial nos níveis de fusão de dados (baixo nível) e de características (médio nível). A pesquisa busca contribuir para o avanço da percepção multissensorial em sistemas ADAS e veículos autônomos, oferecendo propostas compatíveis com arquiteturas modernas de visão computacional e aplicáveis a cenários urbanos complexos.

1.2 Objetivos

1.2.1 Objetivo geral

Desenvolver e avaliar abordagens de fusão de dados em níveis baixo e médio entre sensores câmera e radar automotivo, com o propósito de melhorar a acurácia da detecção de objetos aplicada à percepção do entorno veicular.

1.2.2 Objetivos específicos

- Converter os dados da base de referência para um formato padronizado, compatível com frameworks de detecção de objetos;
- Adaptar a implementação da fusão em nível baixo, baseada em técnicas de geração de propostas a partir de dados de radar, para o framework selecionado;
- Modificar a arquitetura do framework de detecção de objetos baseado em imagens, integrando também os dados de radar na rede de extração de características (*backbone*), viabilizando a fusão em nível médio;
- Desenvolver uma representação espacial dos dados de radar em formato de imagem e implementar a fusão de características no framework escolhido.

2 Fundamentação teórica

Este capítulo apresenta os fundamentos teóricos que embasam o desenvolvimento deste trabalho, com foco na detecção ambiental por meio da fusão sensorial e no uso de redes neurais convolucionais para detecção de objetos. Na Seção 2.1, são introduzidos os Sistemas Avançados de Assistência ao Condutor e os níveis de automação definidos para a condução veicular. A Seção 2.2 discute os principais sensores utilizados na percepção da cena de trânsito em sistemas ADAS baseados em visão, como câmeras, radares e LiDARs, além dos diferentes níveis de fusão de informações multissensoriais. Em seguida, a Seção 2.3 aborda os detectores de objetos baseados em redes neurais convolucionais, com ênfase nos modelos de um e dois estágios mais consolidados na literatura, detalhando suas arquiteturas, mecanismos de detecção e contribuições para o avanço das soluções em visão computacional.

2.1 Sistemas Avançados de Assistência ao Condutor

Os Sistemas Avançados de Assistência ao Condutor tornaram-se indispensáveis nos veículos modernos. Impulsionado pela crescente demanda por mobilidade, o trânsito tornou-se cada vez mais complexo e, portanto, um desafio ainda maior para todos os usuários das rodovias. O objetivo do ADAS é reduzir as consequências de um acidente, prevenir acidentes de trânsito e, num futuro próximo, facilitar a condução totalmente autônoma (ZIEBINSKI et al., 2016).

Esses sistemas têm demonstrado eficácia na redução de acidentes de trânsito ao permitir a detecção antecipada de obstáculos, a emissão de alertas ao condutor e, em alguns casos, a atuação direta sobre os controles do veículo, como frenagem e direção assistida (PARK; YU, W., 2021). Por isso, muitos desses sistemas estão deixando de ser considerados itens exclusivos de veículos de luxo e passando a compor o equipamento padrão em automóveis de menor custo (ZIEBINSKI et al., 2016).

Atualmente, o desenvolvimento dos sistemas ADAS tem se voltado cada vez mais à proteção de usuários vulneráveis nas rodovias, especialmente no contexto de veículos comerciais (OTTO et al., 2012). A seguir, são apresentados alguns exemplos de aplicação desses sistemas, conforme Ziebinski et al. (2016):

- **Monitoramento de ponto cego (*Blind Spot Detection* – BSD):** monitora as áreas laterais próximas ao veículo que não são facilmente visíveis pelo motorista. Sua função é alertar o condutor por meio de um sinal visual, como um ícone no espelho retrovisor lateral, ou por um aviso sonoro, sempre que houver objetos presentes no ponto cego.

- **Alerta de tráfego cruzado traseiro (*Rear Cross Traffic Alert* – RCTA):** auxilia na prevenção de acidentes ao sair de uma vaga em marcha à ré, situação que pode frequentemente resultar em colisões com pedestres ou ciclistas, ocasionando ferimentos graves.
- **Alerta de saída de faixa (*Lane Departure Warning* – LDW):** monitora as marcações laterais da via e detecta quando o veículo está prestes a sair da faixa de rodagem. Ao analisar o movimento da direção, o sistema pode avaliar se a mudança de faixa foi intencional ou não.
- **Assistência à frenagem de emergência (*Emergency Brake Assist* – EBA):** contribui para a segurança ao oferecer suporte ativo à frenagem, incluindo a frenagem automática em situações de risco iminente. Dessa forma, colisões traseiras podem ser evitadas ou, ao menos, ter seus impactos reduzidos devido à menor velocidade e energia de impacto.
- **Controle de cruzeiro adaptativo com função *Stop&Go* (*Adaptive Cruise Control with Stop&Go* – ACC+S&G):** mantém automaticamente a distância em relação ao veículo à frente, mesmo em condições de trânsito com paradas e retomadas. O sistema pode alertar o condutor ou reduzir ativamente a velocidade se a distância se tornar insuficiente, sendo especialmente útil em congestionamentos e situações de tráfego intenso.

A progressiva incorporação de sistemas ADAS nos veículos modernos está diretamente relacionada à evolução dos níveis de automação na condução. Para padronizar essa evolução, em 2014, a SAE International, introduziu o padrão J3016 “Níveis de Automação de Condução” para os consumidores, apresentado na Tabela 1. O padrão J3016 define os seis níveis distintos de automação de direção, começando pelo nível SAE 0, onde o motorista tem total controle do veículo, até o nível SAE 5, onde os veículos podem controlar todos os aspectos das tarefas de direção dinâmica sem intervenção humana (YEONG et al., 2021).

De acordo com a classificação da SAE J3016, a responsabilidade pela condução permanece inteiramente com o condutor humano nos três primeiros níveis, ainda que os sistemas de assistência, como controle de cruzeiro adaptativo e assistência de manutenção de faixa, possam fornecer suporte parcial em determinadas tarefas. A partir do nível 3, o sistema de condução passa a assumir o controle do veículo em cenários específicos, dispensando a atuação do condutor enquanto o sistema estiver ativo. No entanto, no nível 3, ainda é exigida a presença de um condutor habilitado, capaz de retomar o controle quando solicitado. Nos níveis 4 e 5, o próprio sistema do veículo realiza todas as tarefas de direção. No nível 4, isso acontece apenas em situações específicas e previamente definidas. Já no nível 5, o veículo é totalmente autônomo e pode dirigir em qualquer situação, sem precisar da ajuda de um condutor.

Tabela 1 – SAE J3016 – Níveis de Automação da Direção (SAE INTERNATIONAL, 2019)

Categoria	Nível 0	Nível 1	Nível 2	Nível 3	Nível 4	Nível 5
O que o condutor precisa fazer?	Continuar dirigindo sempre que os recursos de assistência estão ativos, mesmo que não precise usar pedais ou volante. Deve supervisionar constantemente os sistemas; devendo frear, manobrar ou acelerar conforme necessário para manter a segurança.			Não precisa assumir o controle da direção quando esses sistemas estão ativos Quando solicitado pelo sistema, deve-se reassumir a direção.	Nenhuma ação será exigida do condutor.	
O que os sistemas fazem?	Apenas fornecem alertas e assistência momentânea.	Assistência na direção **ou** na frenagem/aceleração.	Assistência na direção **e** na frenagem/aceleração simultaneamente.	Dirigem o veículo sob condições específicas.	Dirigem sob condições específicas, sem necessidade de condutor.	Dirigem em **todas** as condições possíveis.
Exemplos de sistemas	Frenagem automática de emergência; Alerta de ponto cego; Alerta de saída de faixa.	Centralização de faixa **ou** Controle de cruzeiro adaptativo	Centralização de faixa **e** Controle de cruzeiro adaptativo	Piloto automático em congestionamento	Táxi autônomo; Pedais/Volante podem ou não estar instalados.	Igual ao nível 4, mas aplicável em qualquer local e condição

Conforme destacado por Dimitrievski et al. (2019), alcançar os níveis mais avançados de automação (níveis 4 e 5) exige não apenas uma integração completa entre hardware e software, mas também o aprimoramento contínuo de algoritmos capazes de realizar a detecção e o acompanhamento preciso de objetos. Esses sistemas devem ser suficientemente robustos para operar com dados ruidosos, lidar com oclusões temporárias, comportamentos imprevisíveis dos agentes no trânsito e eventuais falhas nos sensores. Assim, o avanço rumo à direção totalmente autônoma depende diretamente da superação desses desafios técnicos e da consolidação de soluções confiáveis para a percepção do ambiente.

2.2 Fusão de sensores para detecção ambiental

A combinação de dados provenientes de diferentes sensores, como câmeras, radares e LiDARs, permite explorar informações complementares e redundantes, resultando em maior exatidão, confiabilidade e robustez na percepção do ambiente ao redor do veículo, especialmente em condições adversas. Esse aprimoramento é possibilitado por técnicas de fusão multissensorial, amplamente utilizadas em sistemas ADAS e condução autônoma. Nesta seção, o conteúdo está organizado em quatro partes principais: inicialmente, apresentam-se os sensores câmera, radar e LiDAR, com foco em seus princípios de funcionamento e principais características; por fim, são abordados os três níveis de fusão de informações, com ênfase na integração entre os sensores de câmera e radar.

2.2.1 Câmera

A visão artificial é uma tecnologia popular que tem sido usada há décadas em disciplinas como robótica móvel, vigilância e inspeção industrial. Esta tecnologia oferece capacidades interessantes devido ao baixo custo dos sensores e fornece uma gama de tipos de informação, incluindo espacial (forma, tamanho, distância), dinâmica (objetos em movimento através da análise do deslocamento entre quadros consecutivos) e semântica (análise de forma). As câmeras no mercado oferecem uma ampla gama de configurações em termos de resolução, taxa de quadros, tamanho do sensor e parâmetros ópticos (YEONG et al., 2021). As câmeras podem ser encontradas em versão mono e estéreo, como mostrado nas Figuras 1(a) e 1(d). Há também as câmeras olho de peixe e as câmeras RGB-D, exibidas nas Figuras 1(b) e 1(c).

Figura 1 – Sensores visuais típicos. (a) câmera monocular, (b) câmera olho de peixe, (c) câmera RGB-D, (d) câmera estéreo (LU et al., 2018).

As câmeras fornecem informações ricas sobre a aparência, como contorno, textura, distribuição de cores e outros detalhes visuais, permitindo alcançar desempenho promissor tanto em precisão quanto em velocidade na detecção de objetos (LIU, Y. et al., 2022). Câmeras olho de peixe são uma variante de câmeras monoculares que oferecem amplo ângulo de visão e são atraentes para evitar obstáculos em ambientes complexos, como espaços estreitos e lotados. No entanto, câmeras monoculares e olho de peixe não são capazes de obter mapa de profundidade (LU et al., 2018).

Para obter o mapa de profundidade, existem duas abordagens principais: triangulação e *Time-of-Flight* (ToF). A triangulação pode ser passiva, como na visão estéreo, ou ativa, como em sistemas de luz estruturada, que projetam padrões de luz infravermelha para estimar profundidade a partir da distorção do padrão. As câmeras ToF medem o tempo que a luz leva para ir do emissor ao objeto e retornar ao detector, calculando a profundidade diretamente em circuitos integrados(ZOLLHÖFER et al., 2018).

As câmeras estéreo exploram as diferenças de perspectiva entre duas imagens, permitindo estimar a distância de objetos à frente do veículo em um intervalo típico de 20 a 30 metros. A redundância proporcionada pela segunda câmera aumenta a confiabilidade do sistema (ZIEBINSKI et al., 2016), porém a precisão é fortemente dependente da calibração, o que o torna sensível às condições ambientais, além de implicar maior carga computacional em comparação a outros sensores (ZHU, Y.; WANG, T.; ZHU, S., 2022).

Já câmeras RGB-D utilizam principalmente luz estruturada, como no primeiro Kinect, ou ToF, como no Kinect V2. Funcionalmente, essas abordagens diferem quanto à resiliência à luz de fundo (por exemplo, em aplicações externas), à qualidade dos dados de profundidade e à robustez ao efeito de múltiplos caminhos, em que a luz percorre trajetos indiretos (ZOLLHÖFER et al., 2018).

2.2.2 Radar

O interesse no uso de radar se expandiu nos últimos anos; esses sensores vêm ganhando popularidade por estarem entre os principais componentes de detecção empregados em sistemas ADAS, direção autônoma e aplicações industriais. A tarefa fundamental de um sistema de radar é detectar os alvos em seus arredores e, ao mesmo tempo, estimar seus parâmetros associados (ABDU et al., 2021). Os radares são sensores ativos que transmitem ondas de rádio e analisam os sinais refletidos para determinar a localização e a velocidade dos objetos, ilustrado na Figura 2. Geralmente, representam os objetos detectados como pontos bidimensionais em uma visão superior (*Bird's Eye View* – BEV), fornecendo o ângulo de azimute, a velocidade instantânea e a distância na direção radial (NABATI; QI, 2021).

Figura 2 – Sensor de radar de longo alcance (AG, 2017).

O radar FMCW (*Frequency Modulated Continuous Wave*) é uma tecnologia de detecção amplamente utilizada nos setores automotivo e industrial. Trata-se de um tipo de radar de onda contínua (CW) que transmite sinais com frequência crescente, denominados chirps, geralmente em forma de onda dente de serra (KUMAR; JAYASHANKAR, 2019). Esses radares também funcionam normalmente em frequências de 24 GigaHertz (GHz), 77 GHz e 79 GHz. A frequência GHz corresponde a comprimentos de onda milimétricos; portanto, eles também são chamados de radares de ondas milimétricas (MMW). Existem três classes principais de sistemas de radar automotivo, dependendo da aplicação: SRR (radar de curto alcance), principalmente para assistência de estacionamento e aviso de proximidade de colisão, MRR (radar de médio alcance), principalmente para detecção de ponto cego, prevenção de colisão lateral/traseira e LRR. (Radar de longo alcance) para controle de cruzeiro adaptativo e detecção precoce de colisões (JAHROMI; TULABANDHULA; CETIN, 2019).

2.2.3 LiDAR

LiDAR (*Light Detection and Ranging*) é uma tecnologia de sensoriamento ativo que calcula a distância até um objeto medindo o tempo de ida e volta de um pulso de laser. Para aplicações robóticas e automotivas, utiliza-se um laser NIR de baixa potência, invisível e seguro para o olho humano, com comprimento de onda entre 900 e 1050 nm (ZHOU, 2022). Em sistemas de direção autônoma, sensores LiDAR com 64 ou 128 canais são amplamente empregados para gerar imagens a laser e nuvens de pontos de alta resolução, podendo ser encontrados nas variantes 1D, 2D ou 3D (YEONG et al., 2021). Esses sensores são classificados, com base no método de varredura do feixe de laser, em duas categorias principais: com varredura e sem varredura. Entre os modelos com varredura, há os mecânicos, como os optomecânicos motorizados, e os não mecânicos, como os baseados em sistemas MEMS, que movimentam apenas o feixe, sem deslocamento de componentes ópticos. LiDARs sem partes móveis, como os do tipo Flash e os com matrizes ópticas em fases (OPA), são denominados *solid-state*. Já os baseados em MEMS são classificados como de estado quase sólido (*quasi-solid-state*) (WANG, D.; WATKINS; XIE, H., 2020).

Os LiDARs mecânicos são amplamente utilizados em pesquisa, sendo uma das principais soluções para varredura ambiental de longo alcance. Eles utilizam componentes ópticos avançados e lentes rotativas acionadas por motores elétricos para direcionar os feixes de laser, oferecendo um campo de visão horizontal de até 360°, o que permite a cobertura completa do entorno do veículo (YEONG et al., 2021). Por outro lado, LiDARs de estado sólido (SSL), por eliminarem o uso de partes móveis como lentes rotativas, reduzem o risco de falhas mecânicas. No entanto, apresentam um campo de visão horizontal mais limitado, geralmente de até 120°, quando comparados aos sistemas mecânicos tradicionais (YEONG et al., 2021). Tecnologias como as baseadas em matrizes ópticas em fases (OPA) permitem varredura com acesso aleatório em todo o campo de visão, possibilitando a observação de áreas específicas de interesse e a variação dinâmica da densidade dos feixes. Com isso, é possível realizar uma varredura ampla em baixa resolução e, em seguida, focar objetos de interesse em alta resolução, otimizando a detecção de formas mesmo em longas distâncias (ZHOU, 2022).

2.2.4 Níveis de fusão: Câmera e Radar

Os métodos de fusão de informações multissensoriais são classificados em três categorias, com base nos diferentes níveis de fusão: fusão de informações de baixo nível, fusão de informações de nível médio e fusão de informações de alto nível. Esses correspondem à fusão em nível de dados, fusão em nível de característica e fusão em nível de decisão, conforme proposto pela teoria tradicional de fusão de dados de múltiplas fontes (LIU, Z. et al., 2022).

Na fusão em nível de decisão (alto nível), cada sensor realiza um algoritmo de detecção ou rastreamento separadamente e posteriormente combina o resultado em uma decisão global (YEONG et al., 2021). Os principais métodos de fusão aplicam a teoria bayesiana, a estrutura de filtragem de Kalman e a teoria de Dempster Shafer. Em algumas literaturas, a lista de alvos de detecção de radar foi usada para verificar os resultados da detecção de visão (WEI et al., 2022), exemplificado na Figura 3(a).

As principais vantagens da fusão de alto nível é a menor carga computacional e a redução de recursos de comunicação necessários. Essa abordagem possibilita a padronização da interface para o algoritmo de fusão, eliminando a necessidade de um conhecimento aprofundado dos algoritmos de processamento de sinal subjacentes (YEONG et al., 2021). No entanto, sua principal limitação está na dificuldade de modelar a função de densidade de probabilidade conjunta dos diferentes tipos de informações de detecção, dado que o ruído entre elas é distinto (WEI et al., 2022). Além disso, o ajuste fino dos algoritmos de fusão tende a ter impacto insignificante na precisão ou na latência dos dados (YEONG et al., 2021).

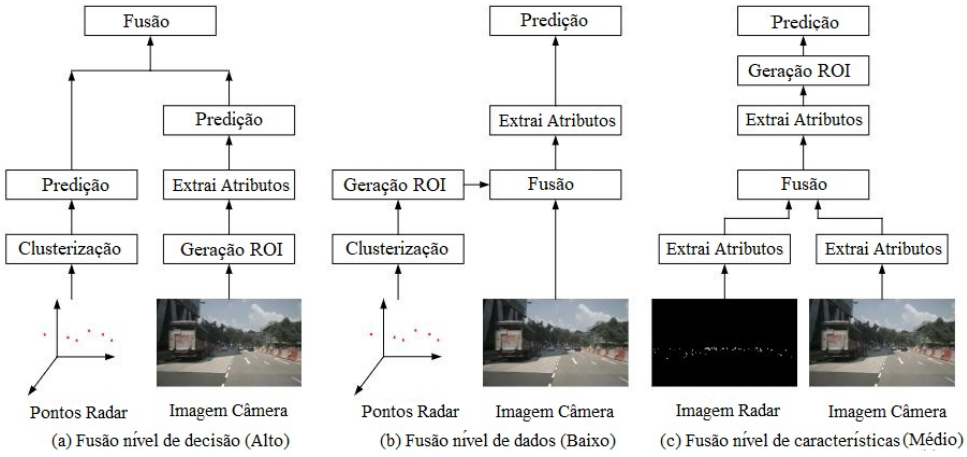

Figura 3 – Arquitetura da fusão radar e câmera por nível (CHANG et al., 2020).

A fusão em nível de dados (baixo nível) gera primeiro a região de interesse (ROI) com base em pontos de radar. A região correspondente da imagem de visão é então extraída de acordo com o ROI. Finalmente, o extrator de recursos e o classificador são usados para realizar a detecção de objetos nessas imagens. Algumas literaturas usam redes neurais para detecção e classificação de objetos (WEI et al., 2022), exemplificado na Figura 3(b).

Esse tipo de fusão permite utilizar as informações captadas por radares antes que as câmeras processem a lista de alvos, o que pode acelerar significativamente os algoritmos de processamento de imagens (WU, X. et al., 2018). Apesar disso, a eficácia da detecção depende diretamente do número de pontos de radar disponíveis. Em casos onde não há pontos radar em determinadas regiões da imagem, essas áreas podem ser ignoradas, comprometendo a segurança (CHANG et al., 2020). Além disso, ao trabalhar com informações em baixo nível, tem-se acesso a uma grande quantidade de dados brutos, o que pode criar desafios relacionados à memória e à largura de banda de comunicação (YEONG et al., 2021).

A fusão em nível de recurso converte os pontos de radar captados no mundo tridimensional (3D) em um plano de imagem bidimensional (2D). As profundidades e velocidades representadas pelos pontos de radar são armazenadas como valores de pixel na imagem transformada. Essa imagem apresenta múltiplos canais, nos quais cada canal corresponde a diferentes estados físicos do ambiente, medidos pelo sensor de radar. Dessa forma, é possível obter dois tipos de representações visuais para a mesma cena de condução: uma imagem de radar e uma imagem da câmera (CHANG et al., 2020), conforme ilustrado na Figura 3(c).

2.3 Detectores de objetos baseados em redes neurais convolucionais

A detecção de objetos é uma tecnologia computacional relacionada à visão computacional e ao processamento de imagens, focada na identificação de instâncias de objetos de uma determinada classe (como humanos, edifícios ou carros) em imagens e vídeos digitais (JIAO et al., 2019). O campo evoluiu consideravelmente com o surgimento das redes neurais convolucionais profundas e ao aumento do poder computacional das GPUs (*Unidade de Processamento Gráfico*). A maioria dos detectores de objetos de última geração utiliza redes de aprendizagem profunda tanto no *backbone*, responsável por extrair características das imagens de entrada, quanto como rede de detecção, que realiza a classificação e a localização dos objetos (JIAO et al., 2019).

A introdução da CNN baseada em região (RCNN) por (GIRSHICK et al., 2013) marcou um avanço significativo, inaugurando uma nova era de progresso para a detecção de objetos. O surgimento dos modelos de detecção de objetos baseados em deep learning trouxe uma distinção clara entre duas abordagens principais: os "detectores de dois estágios"e os "detectores de um estágio"(KHANAM; HUSSAIN et al., 2024). Nos detectores de dois estágios, o primeiro estágio gera propostas de regiões ou objetos, enquanto o segundo estágio classifica essas propostas e ajusta as caixas delimitadoras (SULTANA; SUFIAN; DUTTA, 2020). Em contrapartida, os detectores de um estágio mapeiam diretamente os recursos extraídos para caixas delimitadoras, tratando a tarefa de detecção como um problema de regressão. Embora geralmente mais rápidos, esses detectores tendem a ser menos precisos que os de dois estágios (NABATI; QI, 2020).

A Figura 4 apresenta uma linha do tempo que organiza o lançamento de diferentes detectores de objetos ao longo dos anos. Métodos anteriores a 2012 são classificados como detectores tradicionais, baseados em técnicas clássicas de processamento de imagens. Após esse marco, surgiram os detectores baseados em aprendizado profundo, divididos nas duas categorias principais: os de um estágio, que realizam predições diretamente a partir das características extraídas; e os de dois estágios, que combinam propostas de região com etapas de classificação e refinamento.

No estudo *"An Evaluation of Deep Learning Methods for Small Object Detection"* realizado por Nguyen et al. (2020), conclui-se que os métodos de dois estágios, como o Faster R-CNN (REN et al., 2015), apresentam desempenho superior, demonstrando sua eficácia em diferentes *datasets* e em diversos contextos de detecção de objetos, incluindo aqueles com variação de escalas. Sendo reconhecido como uma referencia (*baseline*) na área, servindo como base para novas pesquisas e desenvolvimentos. Se o objetivo é alcançar um equilíbrio entre precisão e velocidade, o YOLO (REDMON et al., 2016) prova ser uma boa opção, já que o equilíbrio entre velocidade e precisão o torna adequado para aplicações práticas. No entanto, em cenários onde a precisão é priorizada, Faster R-CNN (REN et al., 2015) ou RetinaNet (LIN; GOYAL et al., 2020) continua sendo uma alternativa viável.

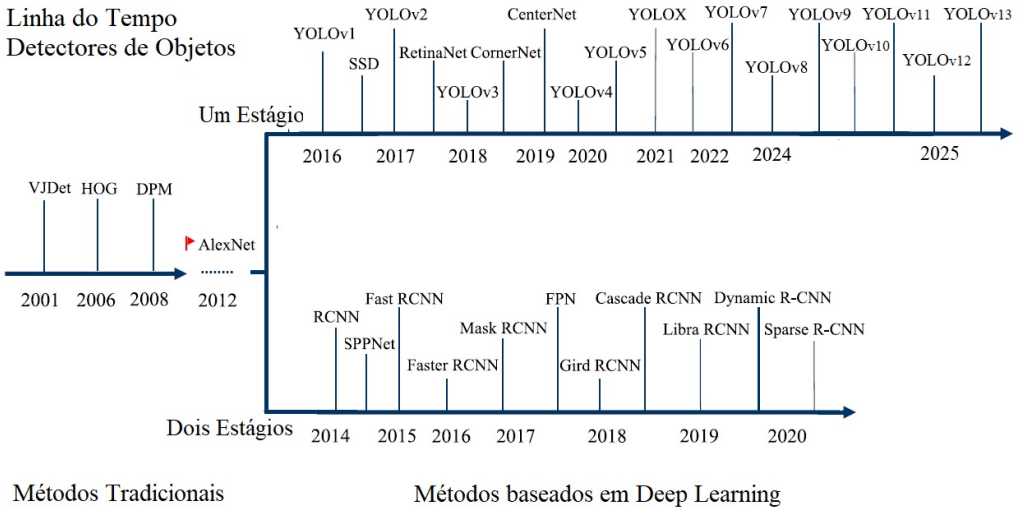

Figura 4 – Linha do tempo da detecção de objetos (LI, Z. et al., 2024).

A seguir, serão abordados os detectores de objetos consolidados na literatura, tais como Fast R-CNN (GIRSHICK, 2015), Faster R-CNN (REN et al., 2015), Mask R-CNN (HE et al., 2017), YOLO (REDMON et al., 2016) e RetinaNet (LIN; GOYAL et al., 2020), com ênfase em suas arquiteturas, propostas e principais contribuições.

2.3.1 Fast R-CNN

Fast R-CNN, proposto por Girshick (2015), é uma extensão do R-CNN que aborda várias de suas limitações, incluindo o treinamento em múltiplas etapas, o custo computacional elevado e o tempo excessivo para detecção de objetos. Essa nova abordagem combina classificação de regiões e regressão de caixas delimitadoras em um único estágio de treinamento, usando uma arquitetura baseada em redes neurais profundas (SULTANA; SUFIAN; DUTTA, 2020).

No R-CNN, cada proposta de região é processada individualmente pela rede convolucional, resultando em cálculos redundantes e altos custos computacionais. O Fast R-CNN resolve esse problema ao processar a imagem inteira, extraindo recursos para todas as regiões de interesse (*RoIs*) uma única vez e enviados à CNN para classificação e localização. Comparado com R-CNN, que insere propostas de cada região para a CNN, uma grande quantidade de tempo para o processamento da CNN e um grande espaço em disco para armazenamento dos recursos pode ser economizado no Fast R-CNN (JIAO et al., 2019).

A arquitetura Fast R-CNN, ilustrada na Figura 5, recebe como entrada uma imagem completa, juntamente com as regiões de interesse (RoIs), que são geradas por métodos externos, como o algoritmo de busca seletiva. A imagem é então processada por uma rede convolucional profunda, como a VGG-16, que extrai um mapa de características convolucionais.

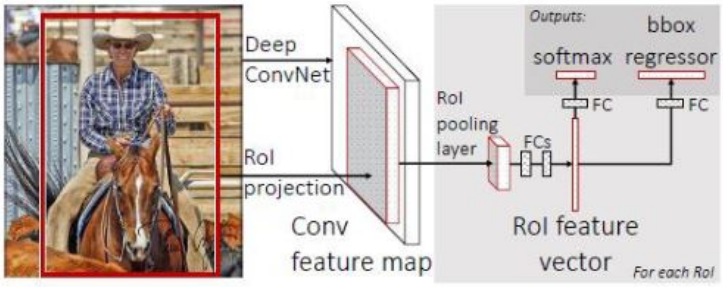

Figura 5 – Arquitetura Fast-RCNN (GIRSHICK, 2015).

Cada RoI é processada por uma camada de *RoI Pooling*, que converte regiões de interesse de tamanhos variados em mapas de características de tamanho fixo. Essa camada divide a região em uma grade uniforme e aplica operações de pooling, como *max pooling*, em cada célula, garantindo uma saída com dimensões consistentes. Posteriormente, cada RoI agrupada é mapeada para um vetor de recursos por camadas *fully connected* (FCs). A rede tem dois vetores de saída por RoI (GIRSHICK, 2015). A primeira camada de saída aplica a função de ativação (*softmax*) para classificar cada região proposta como pertencente a uma das classes de objeto, enquanto a segunda camada realiza a regressão dos quatro parâmetros que definem a caixa delimitadora (*bbox regressor*) associada a cada detecção (JOHN, A.; MEVA, 2020).

Os testes no conjunto de dados PASCAL VOC 2007 demonstraram que Fast R-CNN alcançou um mAP de 66,9% , superando os 66,0% do R-CNN. Além disso, o tempo de treinamento foi reduzido de 84 horas para 9,5 horas, e o tempo de teste por imagem foi reduzido para 0,32 segundos, comparado aos 47 segundos do R-CNN (JIAO et al., 2019).

2.3.2 Faster R-CNN

O Faster R-CNN representa uma evolução significativa em relação ao R-CNN e ao Fast R-CNN ao solucionar a principal limitação relacionada à geração lenta de propostas de regiões (RoIs). O Fast R-CNN ainda depende da busca seletiva, um processo custoso que compromete o desempenho. Conforme Ren et al. (2015), embora o Fast R-CNN atinja taxas quase em tempo real com redes profundas, ignora o tempo gasto na geração de propostas, que permanece como gargalo computacional. A busca seletiva, um dos métodos mais populares, apresenta um tempo médio de execução de cerca de 2 segundos por imagem em implementações baseadas em CPU.

Em contraste, o Faster R-CNN propõe uma solução mais eficiente ao substituir o algoritmo de busca seletiva por uma Rede de Propostas de Região (RPN), uma rede totalmente convolucional (FCN) capaz de gerar propostas com maior rapidez e eficiência, aproveitando as características extraídas pela própria rede de detecção (SULTANA; SUFIAN; DUTTA, 2020). Sua arquitetura, ilustrada na Figura 6, é composta por três partes principais: uma rede de pré-processamento, responsável por extrair características de alto nível da imagem; a RPN, que gera propostas de regiões candidatas; e o cabeçote Faster R-CNN, que refina essas propostas, classificando os objetos e ajustando suas caixas delimitadoras.

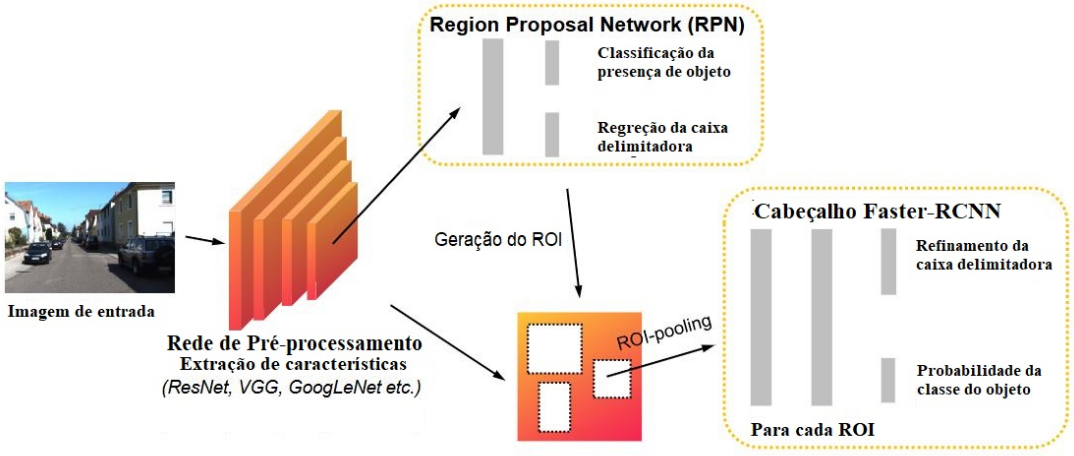

Figura 6 – Arquitetura Faster-RCNN (FENG et al., 2021).

Uma RPN é uma rede totalmente convolucional (FCN) que recebe uma imagem de tamanho arbitrário como entrada e gera um conjunto de propostas de objetos candidatos retangulares. Cada proposta de objeto é associada a uma pontuação de objetividade para detectar se a proposta contém um objeto ou não (SULTANA; SUFIAN; DUTTA, 2020). O RPN é treinado ponta a ponta para gerar propostas de regiões de alta qualidade, que são usadas pelo Fast R-CNN para detecção. Ao Fundir o RPN e Fast R-CNN em uma única rede (Faster R-CNN), compartilhamos seus recursos convolucionais, gerando uma rede neural com mecanismos de atenção (REN et al., 2015).

Os experimentos demonstraram que o Faster R-CNN obteve uma melhoria significativa em termos de precisão e tempo de execução. No conjunto de dados PASCAL VOC 2007, o Faster R-CNN alcançou uma média de precisão (mAP) de 69,9%, superando o Fast R-CNN, que obteve 66,9%, e também apresentou um tempo de execução significativamente menor, reduzindo quase 10 vezes o tempo de processamento de 1830ms para 198ms (JIAO et al., 2019).

2.3.3 Mask R-CNN

O Mask R-CNN avança as técnicas anteriores de detecção de objetos, indo além ao localizar os pixels exatos de cada instância de objeto (segmentação de instância) em vez de apenas delimitar caixas (SULTANA; SUFIAN; DUTTA, 2020). Além de introduzir a segmentação, o Mask R-CNN apresenta maior precisão em relação ao Faster R-CNN devido algumas melhorias na arquitetura. Conforme destacado por (JIAO et al., 2019), o modelo adota a ResNet-FPN como backbone, que combina recursos de múltiplas escalas por meio de uma abordagem de pirâmide de características. Essa estratégia permite a extração de informações semanticamente ricas de mapas de baixa resolução e de detalhes precisos em mapas de alta resolução, sendo especialmente eficaz para detectar objetos pequenos.

A camada *RoI Pooling* também é substituida no modelo pela camada *RoI Align*, que resolve o problema de desalinhamento causado pela quantização nas etapas de pooling. O *RoI Align* utiliza interpolação bilinear para calcular valores exatos em localizações específicas, preservando de forma mais eficiente a informação espacial e garantindo uma correspondência mais precisa entre os RoIs e as características extraídas (JIAO et al., 2019).

Os experimentos mostraram que, com as duas melhorias mencionadas, a precisão foi aprimorada. O uso do backbone ResNet-FPN aumentou em 1,7 pontos a precisão da caixa delimitadora (box AP), enquanto a operação RoI Align contribuiu com um aumento de 1,1 pontos na mesma métrica, no conjunto de dados de detecção MS COCO (JIAO et al., 2019).

2.3.4 YOLO

Em 2016, o campo da detecção de objetos passou por transformações significativas de paradigma com a introdução do *You Only Look Once* (YOLO) por Redmon et al. (2016), um marco que desafiou o paradigma dominante de dois estágios. Ao utilizar uma única rede neural para processar a imagem inteira em uma única passagem, o YOLOv1 apresentou uma abordagem revolucionária que priorizava a velocidade e a simplicidade. Apesar de comprometer a precisão em certos cenários, especialmente para objetos menores, o modelo estabeleceu as bases para futuras iterações, que buscaram equilibrar melhor a relação entre desempenho e eficiência (KHANAM; HUSSAIN et al., 2024).

O pipeline do YOLO trabalha dividindo a imagem de entrada em uma grade de $S \times S$, onde cada célula da grade é responsável por detectar objetos cujo centro está contido nela, conforme ilustrado na Figura 7. O score de confiança é calculado como o produto de duas partes: P(objeto), que representa a probabilidade de a caixa conter um objeto, e o IOU (*Intersection over Union*), que mede a precisão da sobreposição da caixa em relação ao objeto detectado. Cada célula da grade prevê B caixas delimitadoras (x, y, w, h) com seus respectivos escores de confiança, além de probabilidades condicionais de classe em C dimensões para C categorias (JIAO et al., 2019).

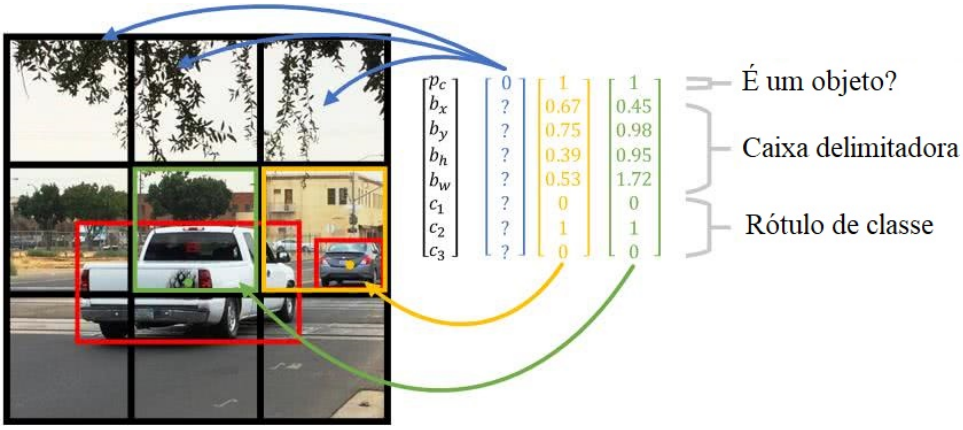

Figura 7 – Pipeline do YOLO (KHARAZI, 2025).

A primeira versão priorizou alta velocidade com uma única CNN, mas teve limitações na precisão, especialmente para objetos pequenos ou sobrepostos. O YOLOv2 introduziu caixas de âncora e camadas de passagem para melhorar a localização dos objetos, enquanto o YOLOv3 trouxe uma arquitetura de extração de características multiescala, aprimorando a detecção em diferentes tamanhos. Nas versões YOLOv4 e YOLOv5, incorporaram backbones otimizados, aumento de dados diversificado e estratégias de treinamento eficientes (TERVEN; CÓRDOVA-ESPARZA; ROMERO-GONZÁLEZ, 2023).

A partir do YOLOv5, os modelos oficiais do YOLO passaram a oferecer escalas ajustáveis para atender a diferentes aplicações e requisitos de hardware (TERVEN; CÓRDOVA-ESPARZA; ROMERO-GONZÁLEZ, 2023). Esses modelos incluem escalas como nano, pequeno, médio, grande e extra-grande. Modelos menores, como *"nano"* e *"pequeno"*, possuem menos parâmetros, sendo mais rápidos e leves, ideais para dispositivos de borda ou aplicações que exigem alta velocidade. Por outro lado, modelos maiores, como *"grande"* e *"extra-grande"*, apresentam maior quantidade de parâmetros, oferecendo maior precisão ao custo de maior demanda computacional. As principais melhorias e características de cada versão do YOLO podem ser acompanhadas na Tabela 2.

Modelo	Ano	Principais Características	Backbone	Conjunto de Dados	mAP	FPS
YOLOv1	2016	Detecção de disparo único, modelo unificado	Darknet-19	PASCAL VOC 2007	63.4%	45
YOLOv2	2017	Caixas de âncora, normalização de lote, recursos refinados	Darknet-19	PASCAL VOC 2007	78.6%	40
YOLOv3	2018	Pontuação de objetividade, previsões multiescala, conexões residuais, classificação multirrótulo	Darknet-53	MS COCO	57.9%	20
YOLOv4	2020	Agregação de características aprimorada, normalização de mini-lotes cruzados (CMBN), aumento de dados diversificado, conexões parciais entre estágios (CSP), ativação Mish	CSPDarknet53	MS COCO	65.7%	33
YOLOv5	2020	Implementação em PyTorch, arquitetura modular, treinamento rápido, design otimizado para diversos hardwares, múltiplos tamanhos de modelo equilibrando velocidade e precisão	EfficientNet-L	MS COCO	55.8%-66.9%	288-140
YOLOv6	2022	Reparametrização, módulos de atenção, cabeça desacoplada	EfficientRep	MS COCO	35.9%-52.5%	802-121
YOLOv7	2022	Backbone e cabeça otimizados, agregação de camadas E-ELAN	E-ELAN	MS COCO	52.8%-73.8%	-
YOLOv8	2023	Detecção sem âncoras, treinamento e inferência mais rápidos, recursos amigáveis ao usuário, camadas convolucionais aprimoradas	CSPDarknet + ConvNeXt	MS COCO	37.3%-53.9%	-
YOLOv9	2024	Introdução de Programmable Gradient Information (PGI) e Generalized Efficient Layer Aggregation Network (GELAN), melhor equilíbrio entre precisão e eficiência	GELAN	MS COCO	38.3%-55.6%	-
YOLOv10	2024	Detecção fim-a-fim sem NMS, dual assignments consistentes, variantes otimizadas para eficiência	CSPNet aprimorado	MS COCO	38.5%-54.4%	543-93
YOLOv11	2024	Blocos C3k2, SPPF otimizado, atenção espacial paralela (C2PSA), suporte a múltiplas tarefas	CSPDarknet-C3k2	MS COCO	39.4%-54.7%	667-88
YOLOv12	2025	Arquitetura centrada em atenção, módulos Area Attention, R-ELAN, suporte a FlashAttention	R-ELAN + Area Attention	MS COCO	40.6%-55.2%	610-85

Tabela 2 – Comparação entre versões do YOLO (KHANAM; HUSSAIN et al., 2024; WANG, C.-Y.; YEH; LIAO, 2024; WANG, A. et al., 2024; KHANAM; HUSSAIN, 2024; TIAN; YE; DOERMANN, 2025).

2.3.5 RetinaNet

Os detectores de objetos de um estágio, embora ofereçam vantagens significativas em termos de velocidade e simplicidade, historicamente apresentam menor precisão quando comparados aos detectores de dois estágios. Essa lacuna de desempenho foi investigada por Lin, Goyal et al. (2020) ao introduzirem o RetinaNet, onde identificaram o desequilíbrio entre as classes de primeiro plano e fundo na fase de treinamento (KHANAM; HUSSAIN et al., 2024).

O RetinaNet é uma rede de detecção de estágio único, unificada e composta por um backbone e duas sub-redes. O backbone, baseado em uma combinação de ResNet e Feature Pyramid Network (FPN), Figura 8(a) e Figura 8(b), é responsável por calcular um mapa de características convolucionais rico e multiescalar a partir de uma imagem de entrada. Sobre esse backbone, são acopladas duas sub-redes: a primeira realiza a classificação convolucional de objetos nas âncoras geradas, Figura 8(c), enquanto a segunda executa a regressão das caixas delimitadoras para ajustá-las às posições e dimensões reais dos objetos detectados, Figura 8(d), (LIN; GOYAL et al., 2020).

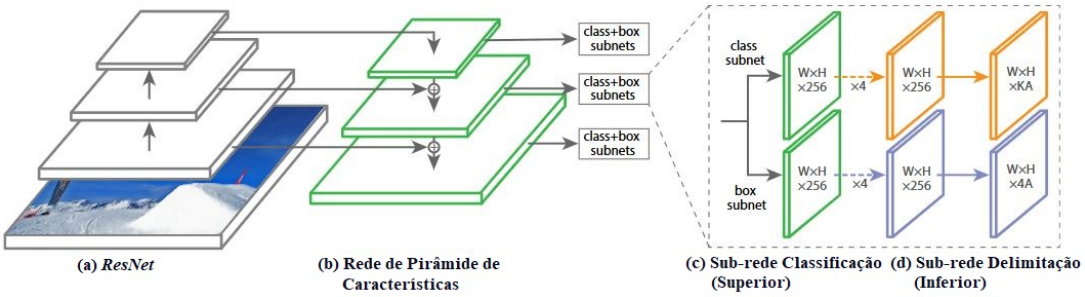

Figura 8 – Arquitetura de rede RetinaNet (LIN; GOYAL et al., 2020).

Conforme autor Lin, Goyal et al. (2020), o design do RetinaNet compartilha várias semelhanças com detectores densos anteriores, especialmente no que diz respeito ao uso de *"âncoras"*, conceito introduzido pela RPN (REN et al., 2015), e à aplicação de pirâmides de recursos, como observado no FPN (LIN; DOLLÁR et al., 2016). No entanto, o grande diferencial do RetinaNet para alcançar resultados superiores em termos de precisão e eficiência não está em inovações no design da rede, mas sim na introdução de uma nova função de perda.

A função de perda, também conhecida como perda focal, é uma perda de entropia cruzada dimensionada dinamicamente, onde o fator de escala decai para zero à medida que a confiança na classe correta aumenta. Intuitivamente, esse fator de escala pode reduzir automaticamente o peso da contribuição de exemplos fáceis durante treinamento e focar rapidamente o modelo em exemplos difíceis (LIN; GOYAL et al., 2020).

Experimentos no conjunto de dados de teste MS COCO, mostram que o RetinaNet com backbone ResNet-101-FPN alcançou um desempenho de 39,1% de AP. Utilizando o ResNeXt-101-FPN, o modelo atingiu 40,8% de AP. Esses resultados destacam a eficiência do RetinaNet em melhorar a precisão da detecção, especialmente para objetos de tamanhos pequenos e médios (JIAO et al., 2019).

3 Estado da arte

O estudo do estado da arte é essencial para identificar avanços, lacunas e tendências em uma área. A revisão da literatura permitiu definir o problema de pesquisa e conhecer técnicas utilizadas. Esta seção está dividida em três partes: metodologia, detalhando o processo de revisão sistemática; revisão sistemática, com os trabalhos relevantes; e técnicas de fusão para detecção 2D, apresentando as estratégias tecnológicas mais promissoras.

3.1 Metodologia

O primeiro passo da metodologia consistiu na formulação das perguntas relacionadas ao tema 'Fusão Sensorial entre Radar e Câmera para Detecção de Objetos no Ambiente de Veículos Autônomos', as quais são apresentadas a seguir:

1. O que é a fusão sensorial e quais são suas aplicações em sistemas de condução assistida?
2. Quais são os benefícios e limitações dos principais sensores utilizados na percepção da cena de trânsito?
3. Quais são os principais desafios técnicos na integração de dados de radar e câmera em redes neurais profundas?
4. Quais implementações de fusão entre radar e câmera têm demonstrado melhor desempenho em redes neurais profundas aplicadas à percepção da cena de trânsito?

A fim de obter as palavras-chave mais relevantes, realizou-se uma pesquisa na base de dado Scopus com a seguinte query: 'sensor fusion' E 'camera' E 'radar', com a janela temporal de 2020 a 2022. Dos 712 artigos obtidos, foram selecionados os 100 mais relevantes, os quais foram posteriormente submetidos ao VOSViewer para uma análise das palavras encontradas nos resumos. A partir dessa análise, foram extraídas 54 palavras que apareceram no mínimo 8 vezes nesse contexto. Dentre elas, as palavras-chave selecionadas para nossa pesquisa são apresentadas na tabela 3:

Tabela 3 – Palavras-Chave escolhidas

Palavra-Chave	Termos Associados
KW1	"Sensor Fusion", "multi-sensor", "data fusion"
KW2	"Camera"
KW3	"Radar"
KW4	"Autonomous Vehicles", "Autonomous Driving", "self-driving", "autonomous car"
KW5	"Advanced Driver Assistance Systems", "ADAS"
KW6	"Object Detection", "Object Tracking", "Semantic Segmentation"

Com base na combinação das perguntas de pesquisa e palavras-chave selecionadas, a seguinte query foi elaborada: "Sensor Fusion"OR "multi-sensor"OR "data fusion"AND "camera"AND "radar"AND ("ADAS"OR "Advanced driver-assistance"OR "Autonomous Vehicles"OR "Autonomous Driving"OR "self-driving"OR "autonomous car"). Essa consulta foi utilizada para realizar buscas nas bases de dados IEEE Xplore, Scopus e Tufts JumboSearch, resultando inicialmente em 650 artigos. Após a remoção de duplicatas, o número final de artigos selecionados foi reduzido para 425.

Tabela 4 – Processo de triagem de artigos

Etapa	AÇÃO	Y	Total de artigos
0	Combinar 3 bases de dados (IEEExplore, Tufts JumboSearch, Scopus)		650
0	Informações duplicadas e irrelevantes removidas		225
0	Nova população	Y0	425
	TRIAGEM AUTOMÁTICA DE TÍTULOS		
-	População	Y0	425
1	Triagem automática de títulos (KW1+KW2+KW3+KW6 $\geq$ 3)	Y1	33
-	Nova população	Y0-Y1	392
	TRIAGEM MANUAL DE TÍTULOS		
-	População	Y0-Y1	392
1	Pesquisador / Orientador (Y&Y)	Y2	21
1	Pesquisador / Orientador (Y&M OU M&Y)	Y3	32
1	Pesquisador / Orientador (Y&N OU M&M OU N&Y)	Y4	146
1	Pesquisador / Orientador (N&N OU M&N OU N&M)		193
-	Nova população (Y1+Y2+Y3+Y4)	Y5	232
	TRIAGEM AUTOMÁTICA DE RESUMOS		
-	População	Y5	232
2	Palavras-chave de pesquisa (KW1+KW2+KW3+KW5+KW6 = 5)	Y6	15
2	Palavras-chave de pesquisa (KW1+KW2+KW3+KW5+KW6 < 5)		217
-	Nova população	Y5-Y6	217
	TRIAGEM MANUAL DE RESUMOS		
-	População	Y5-Y6	217
2	Pesquisador / Orientador (Y&Y)	Y7	48
2	Pesquisador / Orientador (Y&M OU M&Y OU N&Y OU Y&N)	Y8	37
2	Pesquisador / Orientador (N&N OU M&N OU N&M)		132
-	Nova população (Y6+Y7+Y8)	Y9	100
	TRIAGEM DO TEXTO COMPLETO		
-	População	Y9	100
3	Não encontrado ou sem acesso		15
3	Critérios de inclusão (Contribuição $\geq$ 2; Teoria $\geq$ 2; Metodologia $\geq$ 1; Análise de dados = TODOS)		42
3	Critérios de exclusão (Contribuição $\leq$ 1; Teoria $\leq$ 1)		43
	RESULTADOS		
	TOTAL DE ARTIGOS INCLUÍDOS		42
	TOTAL DE ARTIGOS EXCLUÍDOS		608
	PERCENTUAL DE ARTIGOS INCLUÍDOS		7%

A triagem dos artigos foi realizada em três etapas. Na primeira, os títulos foram filtrados automaticamente, aprovando aqueles que continham ao menos três das quatro palavras-chave KW1, KW2, KW3 e KW6. Os demais foram avaliados manualmente, com leitura dos títulos e validação cruzada entre pesquisador e orientador; artigos marcados como *negado por ambos* (N&N) ou *talvez e negado* (M&N ou N&M) foram recusados. Como resultado, 33 artigos foram aprovados automaticamente e 199 por triagem manual, totalizando 232 para a próxima etapa (Tabela 4). Na segunda etapa, os resumos foram analisados: aqueles que continham simultaneamente as palavras-chave KW1, KW2, KW3, KW5 e KW6 foram automaticamente aprovados, enquanto os demais passaram por triagem manual com validação cruzada, resultando na exclusão de 132 artigos e aprovação de 100. Na etapa final, os textos completos foram avaliados superficialmente com notas de 0 a 3 para contribuição, teoria e metodologia. Apenas os que obtiveram pelo menos 2 em contribuição e teoria foram considerados elegíveis; artigos inacessíveis ou fora dos critérios também foram excluídos. Ao final, 42 artigos foram selecionados para leitura detalhada.

3.2 Análise temática

Nesta seção, o objetivo é responder às perguntas de pesquisa com base na literatura disponível sobre a fusão sensorial entre radar e câmeras para detecção de objetos no ambiente automotivo. As perguntas, formuladas na seção de planejamento, serão abordadas a partir da análise dos artigos selecionados. A revisão da literatura será organizada em torno de temas principais, que agruparão as questões a serem respondidas. Para a elaboração desta etapa, foi utilizado o software ©Nvivo, que permitiu realizar o agrupamento (clusterização) das informações extraídas dos artigos, como um esforço do pesquisador para responder às perguntas de pesquisa com base na análise crítica da literatura.

Quais são os requisitos necessários para aplicar a fusão entre os sensores câmeras e radar para tarefa de detecção de objetos?

- **Tema 1: Fusão sensorial e sua aplicação voltada para sistemas automotivos.**
  O que é a fusão sensorial e qual sua aplicação no ambiente de condução assistida?
- **Tema 2: Benefícios e limitações dos sensores câmera, radar e LiDAR para detecção de objetos no contexto automotivo.**
  Quais os benefícios e limitações entre os principais sensores utilizados na percepção da cena de trânsito?
- **Tema 3: Oportunidades e desafios na aplicação de redes neuronais multissensoriais**
  Quais são os principais desafios técnicos na integração de dados de radar e câmera em redes neuronais profundas?

Para responder ao último questionamento, este estudo adota a tarefa de detecção de objetos 2D como foco principal de análise. Na Seção 3.3, são apresentadas as principais arquiteturas de redes neurais propostas na literatura para a fusão de dados de radar e câmera, acompanhadas de suas respectivas classificações com base nas métricas mais utilizadas para avaliação do desempenho em detecção de objetos 2D.

3.2.1 Tema 1: Fusão sensorial e sua aplicação voltada para sistemas automotivos.

A fusão de sensores tem suas origens em aplicações militares, onde a integração de informações provenientes de diversas fontes foi empregada para desenvolver uma visão mais completa e precisa de campos de batalha ou situações de combate(ALTENDORFER; WIRKERT; HEINRICHS-BARTSCHER, 2010). O objetivo principal das aplicações de fusão de dados é combinar informações de sensores individuais de maneira que seus pontos fortes sejam maximizados e suas limitações minimizadas. Tipicamente, as configurações de fusão de dados abordam aspectos como redundância e complementaridade da informação, aprimoramento da temporalidade dos dados e redução de custos (DARMS et al., 2010).

De acordo com Altendorfer, Wirkert e Heinrichs-Bartscher (2010), a fusão de sensores oferece uma série de vantagens gerais que tornam os sistemas mais eficazes e confiáveis em aplicações complexas. Entre os principais benefícios, destacam-se:

- **Robustez:** A redundância proporcionada pelo uso de múltiplos sensores aumenta a resistência do sistema a falhas parciais, garantindo maior confiabilidade em condições adversas.
- **Cobertura ampliada:** Quando os alcances de diferentes sensores não se sobrepõem ou apresentam apenas sobreposições parciais, a fusão sensorial permite expandir significativamente a cobertura conjunta.
- **Maior confiança:** As medições realizadas por um sensor podem ser confirmadas por outras fontes sensoriais que monitoram o mesmo domínio, elevando a confiabilidade dos dados coletados.
- **Melhoria na precisão:** A combinação de dados provenientes de múltiplos sensores que monitoram o mesmo domínio permite medições mais precisas de grandezas como distância, velocidade e outras variáveis relevantes.

Com o aumento da complexidade do tráfego, os Sistemas Avançados de Assistência ao Condutor tornaram-se essenciais nos veículos modernos, visando reduzir as consequências de acidentes, prevenir colisões e, no futuro, possibilitar a condução totalmente autônoma (ZIEBINSKI et al., 2016). Nesse contexto, a abordagem tradicional para direção autônoma tem passado por mudanças significativas. Em vez de se basear em um único tipo de sensor, como câmeras, radares ou LiDAR, as soluções atuais integram diferentes sensores para criar sistemas mais robustos e adaptáveis. Essa combinação permite melhorar o desempenho em condições variadas, ao mesmo tempo que considera a viabilidade econômica dessas tecnologias (KUMAR; JAYASHANKAR, 2019). Um exemplo dessa transformação é apresentado por Wei et al. (2022), que analisaram os dados de grandes fabricantes sobre o uso de sensores em veículos autônomos. A Tabela 5, fundamentada neste estudo, ilustra como as principais montadoras estruturaram suas soluções, destacando variações tanto na quantidade quanto nos tipos de sensores empregados.

Empresa	Sistema de Direção	Configuração dos Sensores
Tesla	Autopilot	8 câmeras, 12 radares ultrassônicos, radar mmWave
Baidu	Apollo	Lidar, radar mmWave, câmera
NIO	Aquila	Lidar, 11 câmeras, 5 radares mmWave, 12 radares ultrassônicos
Xpeng	XPILOT	6 câmeras, 2 radares mmWave, 12 radares ultrassônicos
Audi	Traffic Jam Pilot	6 câmeras, 5 radares mmWave, 12 radares ultrassônicos, Lidar
Mercedes Benz	Drive Pilot	4 câmeras panorâmicas, Lidar, radar mmWave

Tabela 5 – Soluções de sensores de direção autônoma de alguns fabricantes (WEI et al., 2022).

Em sistemas de direção autônoma ou ADAS, vários sensores são frequentemente usados para melhorar a redundância e a tolerância a falhas do sistema. Como a função de detecção não pode ser alcançada por um único sensor, o objetivo da fusão de dados multissensor é usar informações redundantes e informações complementares fornecidas por vários números ou tipos de sensores para reduzir a incerteza e a ambiguidade das informações de observação e aumentar a confiabilidade e a capacidade de sobrevivência do sistema de detecção (LIU, Z. et al., 2022).

Descoberta: A fusão sensorial desempenha um papel crucial no ambiente de veículos autônomos, integrando dados de diferentes sensores para melhorar a percepção do entorno e a tomada de decisão. Essa abordagem permite combinar informações complementares e redundantes, aumentando a precisão, confiabilidade e robustez dos sistemas de detecção em condições adversas. Além disso, a fusão sensorial é essencial para superar as limitações de sensores individuais, como a incapacidade de lidar isoladamente com diferentes condições ambientais, tornando-se uma tecnologia indispensável para alcançar a segurança e a eficiência necessárias na condução autônoma.

3.2.2 Tema 2: Benefícios e limitações dos sensores câmera, radar e LiDAR para detecção de objetos no contexto automotivo.

Os principais sensores utilizados para percepção ambiental em veículos autônomos são a câmera, o radar e o LiDAR. Cada um desses sensores apresenta benefícios e limitações próprios, conforme ilustrado na Figura 9.

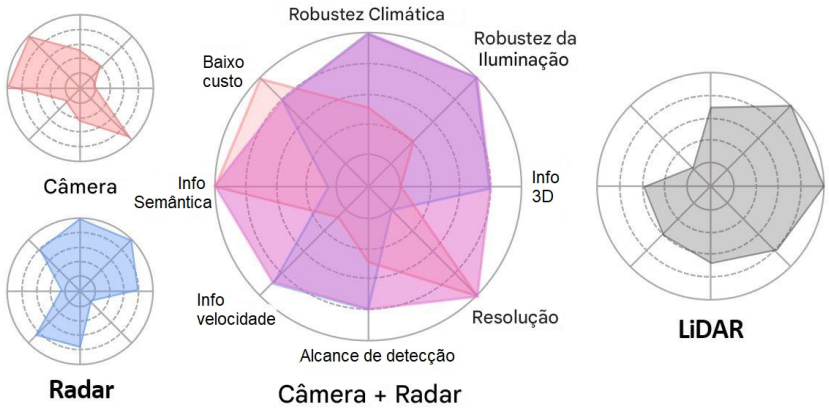

Figura 9 – Vantagens e limitações dos principais sensores utilizados para percepção ambiental (KIM, Y. et al., 2022).

Os avanços nas pesquisas com detectores de objetos baseados em redes neurais convolucionais (CNNs) têm proporcionado alta precisão em condições favoráveis, como dias ensolarados e ambientes com boa iluminação (LI, L. Q.; XIE, Y. L., 2020). No entanto, esses algoritmos ainda enfrentam desafios relevantes em cenários reais, caracterizados pela diversidade de objetos, incluindo pedestres, carros, caminhões, bicicletas e motocicletas, que apresentam variações de escala e proporção. Além disso, o desempenho da detecção é significativamente comprometido por oclusões parciais e por condições climáticas adversas, como chuva intensa e neblina (MICHAELIS et al., 2019).

Embora a fusão entre câmera e LiDAR tenha alcançado uma boa taxa de detecção de alvos, atender aos requisitos em tempo real continua sendo um desafio, devido ao grande volume de informações adquiridas e ao alto custo computacional envolvido. Além disso, o desempenho de reconhecimento tanto das câmeras quanto dos sensores LiDAR é significativamente afetado por condições climáticas adversas, limitando a robustez do sistema em ambientes de tráfego complexos (JIANG; ZHANG, L.; MENG, 2019). Por outro lado, os radares automotivos oferecem vantagens importantes, como ampla capacidade de adaptação a diferentes condições ambientais, penetração eficaz em chuva e nevoeiro, além de fornecer diretamente informações sobre profundidade, velocidade e um amplo alcance de detecção (CHAVEZ-GARCIA et al., 2012).

Comparados aos LiDARs, os radares oferecem maior alcance de detecção e uma certa capacidade de penetração, o que os torna mais adequados para enfrentar condições desafiadoras. Além disso, os radares são mais econômicos na prática, favorecendo seu uso em detrimento dos LiDARs (LIU, Y. et al., 2022). Já que para a produção em massa, o custo geralmente é o principal critério na escolha do sensor, tornando o uso do LiDAR raro, apesar de seu alto desempenho (KANG; KUM, 2020).

Apesar da sua robustez, o radar automotivo também apresenta algumas limitações importantes. Os pontos de radar são significativamente esparsos, o que dificulta a estimativa de informações geométricas, como localização e dimensões, além de comprometer a classificação precisa dos objetos (LIU, Y. et al., 2022). Apesar de sua capacidade de detectar objetos a longas distâncias, os sinais retornados frequentemente contêm ruídos provenientes do solo, edificações e grades, o que contribui para o aumento de *Falsos Positivos* (FP), especialmente em pequenos alvos(LI, L. Q.; XIE, Y. L., 2020). A identificação de pedestres é outro desafio, pois sua seção transversal de radar (*Radar Cross Section, RCS*) é consideravelmente menor em comparação a outros usuários da cena de transito, tornando sua detecção difícil em cenas desordenadas, principalmente quando estão estáticos, parcialmente obstruídos ou próximos a objetos altamente refletivos, como veículos, postes, semáforos e placas de sinalização (DIMITRIEVSKI et al., 2019).

Descoberta: A combinação de câmeras e radares é uma solução eficiente e econômica. As câmeras, com sua capacidade de detectar objetos utilizando redes neurais convolucionais (CNNs), oferecem informações detalhadas sobre a aparência e o contexto visual, desempenhando um papel crucial na detecção de pedestres, superando as limitações dos radares, que têm dificuldade em identificar pedestres devido à baixa seção transversal de radar (RCS). Por outro lado, os radares são extremamente robustos em condições climáticas adversas, como chuva, nevoeiro e baixa iluminação, onde as câmeras apresentam desempenho reduzido. Além disso, os radares fornecem dados confiáveis de profundidade e velocidade, complementando as informações visuais das câmeras. Ambos os sensores são econômicos e amplamente utilizados na indústria automotiva, facilitando sua integração em veículos de produção em massa. Assim, a fusão entre câmeras e radares oferece uma solução balanceada e eficaz, combinando precisão, robustez e viabilidade econômica.

3.2.3 Tema 3: Oportunidades e desafios na aplicação de redes neuronais multissensoriais

O melhor desempenho alcançado pelas redes neuronais no processamento de dados baseados em imagens fez com que os pesquisadores incorporassem modalidades de detecção adicionais na forma de fusão de sensores. Com esse objetivo, os modelos de aprendizagem profunda estão sendo expandidos para realizar a fusão multissensor profunda, a fim de se beneficiarem dos dados de complementaridade de modelos de detecção múltipla, particularmente em situações ambientais complexas, como no caso da condução autônoma (ABDU et al., 2021).

Diferentemente do aprendizado de máquina tradicional, que exige a engenharia manual de recursos para extrair características relevantes, o aprendizado profundo automatiza essa etapa. Para isso, depende de hardware avançado, como GPUs, capazes de otimizar operações complexas, como a multiplicação de matrizes, que são fundamentais para o treinamento de modelos com grande número de parâmetros. Essa capacidade permite lidar com tarefas de maior complexidade, embora exija um poder computacional significativamente superior aos métodos clássicos. Além disso, como requer grandes quantidades de informação para um treinamento eficaz, o aprendizado profundo apresenta desempenho superior em cenários com grandes volumes de dados, mas sua eficácia tende a diminuir quando os conjuntos de dados são reduzidos (ABDU et al., 2021).

Segundo Park e Wenchang Yu (2021), é uma tarefa desafiadora desenvolver um sistema de classificação de objetos com um conjunto de dados relativamente pequeno. Em geral, a uma rede neural treinada com um pequeno número de amostras de dados é propensa a baixo desempenho e overfitting. No entanto, ao utilizar modelos CNN de última geração treinados com grandes volumes de dados, os recursos aprendidos podem ser reaproveitados em um novo sistema com um conjunto de dados menor. Esse processo, conhecido como aprendizado de transferência, permite que modelos CNN pré-treinados sejam usados como ponto de partida para novas tarefas. Essa abordagem proporciona uma precisão geral superior em comparação ao treinamento de um modelo a partir do zero.

Os sinais de radar enfrentam desafios consideráveis em sua aplicação com algoritmos de aprendizado profundo, principalmente devido à escassez de conjuntos de dados de acesso público e à ausência de anotações de objetos. Como consequência, muitos pesquisadores desenvolvem seus próprios conjuntos de dados para avaliar e validar os modelos propostos. No entanto, a criação desses conjuntos é um processo demorado, especialmente quando se busca alcançar uma escala adequada. Frequentemente, as bases de dados geradas e seus benchmarks não são disponibilizados publicamente, o que dificulta a comparação entre algoritmos e limita o avanço das pesquisas baseadas em sinal de radar com redes neurais (ABDU et al., 2021).

A qualidade e o alinhamento dos dados também são fatores críticos: o sistema de visão deve ser calibrado tanto espacial quanto temporalmente, pois desalinhamentos durante a coleta dos dados de treinamento podem introduzir erros significativos nos conjuntos, comprometendo o desempenho das redes neurais (FENG et al., 2021). A calibração é fundamental para etapas posteriores do processamento de dados, como a fusão sensorial, a detecção de obstáculos, a localização, o mapeamento e o controle do veículo (YAN et al., 2022). No contexto da percepção em tempo real, as informações ambientais captadas em momentos distintos podem apresentar discrepâncias significativas, especialmente devido ao movimento do veículo e às variações no ambiente. Por isso, é essencial que os dados obtidos por diferentes sensores estejam sincronizados no tempo, de modo a permitir uma fusão eficaz das informações (ZHANG, X. et al., 2019). Existem duas abordagens principais para a calibração temporal dos sensores: a sincronização externa, que utiliza hardware dedicado para alinhar os tempos de aquisição, e a sincronização interna, que explora os carimbos de data e hora gerados por cada sensor para realizar o alinhamento temporal (YEONG et al., 2021). Em paralelo, a calibração espacial entre sensores, como radar e câmera, é frequentemente abordada na literatura como calibração de coordenadas, cujo objetivo é alinhar os pontos do radar com os objetos detectados nas imagens. Para isso, os métodos mais utilizados são classificados em três categorias principais: transformação de coordenadas, verificação entre sensores e abordagens baseadas em visão (WEI et al., 2022).

- Método de transformação de coordenadas: O método de transformação de coordenadas unifica as informações de radar e de visão sob o mesmo sistema de coordenadas através de operações matriciais;
- Método de verificação de sensor: O método de verificação de sensor calibra vários sensores entre si com as informações de detecção de diferentes sensores no mesmo objeto. Primeiro, a lista de alvos é gerada pelo radar e depois a lista é verificada pelas informações de visão;
- Método baseado em visão: Utiliza de técnicas como subtração adaptativa de fundo ou movimento estéreo para achar a correspondência de objetos de radar e objetos de imagem.

Atualmente, há uma maior disponibilidade de conjuntos de dados públicos que incluem informações de radar. A Tabela 6, adaptada de Sheeny et al. (2021), apresenta uma comparação entre os principais conjuntos de dados automotivos, destacando aspectos como os sensores utilizados, as condições ambientais contempladas e os tipos de anotações oferecidas.

Tabela 6 – Conjuntos de dados automotivos públicos com detecção de radar (SHEENY et al., 2021)

Base	Tam	Radar	LiDAR	Câmera	Noite	Nevoa	Chuva	Neve	Detec Obj	Rastre Obj	Odo metria	Anot 3D
nuScenes	G	Nuvem de pontos esparsa	✓	✓	✓		✓		✓	✓		✓
Oxford Radar RobotCar	G	Imagem de radar de alta resolução	✓	✓	✓		✓				✓	
MulRan	G	Imagem de radar de alta resolução	✓								✓	
Astyx	P	Nuvem de pontos esparsa	✓	✓					✓			✓
RADIATE	G	Imagem de radar de alta resolução	✓	✓	✓	✓	✓	✓	✓	✓	✓	Pseudo-3D

A base nuScenes (CAESAR et al., 2020) oferece dados de radar em formato de nuvem de pontos esparsa, juntamente com LiDAR e câmeras, cobrindo condições climáticas como noite e chuva, além de incluir anotações para detecção e rastreamento de objetos. Já o Oxford Radar RobotCar (BARNES et al., 2019) e o MulRan (KIM, G. et al., 2020) fornecem imagens de radar de alta resolução, mas focam principalmente em aplicações de localização e mapeamento, sem anotações específicas para objetos. O Astyx (MEYER, 2019), embora seja um conjunto de dados pequeno, com cerca de 500 quadros anotados, inclui anotações 3D. O RADIATE (SHEENY et al., 2021) oferece imagens de radar de alta resolução, juntamente com LiDAR e câmeras, abrangendo condições ambientais adversas, como neblina, chuva e neve.

Descoberta: As redes neurais multissensor emergem como a chave para aumentar a robustez e a precisão dos modelos que utilizam apenas câmeras. A transferência de aprendizado é uma solução valiosa para aproveitar modelos já treinados, que possuem alta eficiência em tarefas de detecção, permitindo a adaptação desses modelos para novos contextos e proporcionando uma precisão superior à de um treinamento realizado completamente do zero. No entanto, a criação de uma nova base de dados para treinar modelos multissensor é uma tarefa custosa e demorada, que exige anotações detalhadas de todos os objetos de interesse na cena, além da calibração adequada entre os sensores, tanto espacial quanto temporalmente, para garantir a precisão dos dados. Felizmente, já existem bases de dados públicas que atendem a esses requisitos, como nuScenes, Oxford Radar RobotCar, MulRan e RADIATE, que fornecem dados de radar, câmeras e LiDAR em condições variadas, com anotações para tarefas como detecção, rastreamento de objetos e odometria. A utilização dessas bases de dados públicas facilita o desenvolvimento e aprimoramento de modelos de aprendizado profundo, pois permite a comparação de algoritmos e a validação de novos métodos, acelerando o progresso na área.

3.3 Técnicas de fusão entre câmera e radar para detecção 2D

A área de visão computacional abrange uma ampla gama de tarefas, incluindo segmentação semântica, detecção e rastreamento de objetos em 2D e 3D, entre outras. Cada uma dessas atividades envolve técnicas específicas, com desafios e limitações que exigem soluções adaptadas às suas particularidades. Dentre elas, a detecção de objetos se destaca como um tema central, especialmente devido à sua relevância para a navegação de veículos autônomos. Com o objetivo de proporcionar uma compreensão abrangente das abordagens descritas na literatura, esta seção se concentra na análise das principais técnicas de fusão sensorial que utilizam sinais de radar e câmera para a detecção de objetos em 2D.

Na Tabela 8, apresenta-se um resumo detalhado das técnicas exploradas em 14 artigos distintos, abordando as arquiteturas utilizadas, o nível de fusão adotado, a operação de fusão empregada, o problema investigado, os tipos de objetos identificados e as bases de dados utilizadas. Complementarmente, a Tabela 7 reúne os resultados quantitativos reportados com base na base de dados nuScenes, utilizando métricas como Precisão Média (AP) e Revocação Média (AR). A análise dessas tabelas permite comparar as diferentes arquiteturas propostas, identificar padrões de projeto e destacar as abordagens com desempenho mais robusto na detecção de objetos em 2D, fornecendo uma base sólida para decisões de implementação e futuras otimizações.

Alguns estudos também reportaram a latência dos algoritmos como uma métrica de desempenho. No entanto, esses valores não foram incluídos na análise comparativa apresentada, pois a latência não pode ser diretamente comparada entre trabalhos que utilizam hardwares distintos, o que compromete a validade da comparação. De forma semelhante, também foram desconsideradas métricas obtidas em bases distintas da nuScenes, devido às variações que diferentes conjuntos de dados podem introduzir nos resultados.

Tabela 7 – Quadro resumo das métricas obtidas por trabalhos de fusão sensorial entre RADAR e Câmera para detecção de Objetos 2D na base de dados NuScenes

Reference	Scale	AP	AP^{50}	AP^{75}	mAP	AR	Repository
Chang et al. (2020)	800	72.4	90.0	79.3	-	79.0	SAF-FCOS
Yadav, Vierling e Berns (2020)	1024	72.3	88.9	84.3	-	75.3	BIRANet
Yadav, Vierling e Berns (2020)	512	68.7	87.6	79.7	-	72.0	BIRANet
Yadav, Vierling e Berns (2020)	512	64.7	82.1	57.4	-	67.5	RANet
Nabati e Qi (2019)	-	35.4	59.0	37.4	-	42.1	RRPN
Nobis et al. (2020)	640	-	-	-	43.9	-	CRF-Net
V. John, Nithilan et al. (2020)	224	42.3	-	-	-	-	SO-Net
Liang Qun Li e Yuan Liang Xie (2020)	800	24.3	48.4	22.3	-	33.7	Li-Xie
Nabati e Qi (2020)	-	35.6	60.5	37.4	44.5	42.1	Nabati-Qi
Vijay John e Mita (2019)	416	56.0	-	-	-	-	RVNet

Tabela 8 – Quadro resumo das técnicas utilizadas pelos pesquisadores para fusão sensorial entre câmera e radar automotivo

Reference	Network Architecture	Level of Fusion	Fusion Operation	Problem	Object Type	Data set
Chang et al. (2020)	SAF-FCOS based on FCOS	Feature level	Addition; Multiplication	2D Object detection	Bicycle, car, motorcycle, bus, train, truck	NuScenes
Yadav, Vierling e Berns (2020)	BIRANet based on ResNet	Feature level	Addition	2D Object detection and distance estimation	Car, Truck, Person, Motorcycle, Bicycle, Bus	NuScenes
Nabati e Qi (2019)	RRPN based Fast-R-CNN	Data level	Transformation matrix	2D Object detection	Car, Truck, Person, Motorcycle, Bicycle and Bus	NuScenes
Nobis et al. (2020)	CRF-Net based RetinaNet with VGG	Multi-level	Feature concatenated	2D Object detection	Car, bus, motorcycle, truck, trailer, bicycle and human	NuScenes
V. John, Nithilan et al. (2020)	SO-Net based Yolov3 and Encoder-decoder	Feature level	Concatenation	2D Object detection and Free space Segmentation	Vehicles and free space	NuScenes
Liang Qun Li e Yuan Liang Xie (2020)	Li-Xie based on the YOLOv3	Feature level	Concatenation; Multiplication	2D Object detection	car, bus, truck, trailer	NuScenes
Nabati e Qi (2020)	Nabati-Qi based on Fast-R-CNN	Multi-level	Region proposal	2D Object detection and distance estimation	Car, Truck, Person, Bus, Bicycle, Motorcycle	NuScenes
Vijay John e Mita (2019)	RVNet based on YOLOv3	Feature level	Concatenation	2D Object detection	vehicles, pedestrians, two-wheelers and objects (moving and debris)	NuScenes
Kang e Kum (2020)	VGG16	data level	Transformation matrix	Vehicle localization	vehicles	own, Stanford
Park e Wenchang Yu (2021)	VGG-19, GoogLeNet e VGG-16	data level	Transformation matrix	2D Object detection	bicycle, car and pedestrian	Udacity vehicle, INRIA Person and others
Ze Liu et al. (2022)	Based Faster R-CNN	decision level	JPDA	Target Recognition and Tracking	Vehicles and pedestrians	Own, MS COCO 2014, VOC2007 and VOC2012
Xinyu Zhang et al. (2019)	RCNN	data level	Transformation matrix	2D Object detection	car, trucks and vans	Own
Han et al. (2016)	Model based machine learning (DPM)	data level	Transformation matrix	2D Object detection	Vehicles, Pedestrians, Two Wheels, Traffic Cones	PASCAL VOC2010, INRIA Person
Jiang, Lijun Zhang e Meng (2019)	YOLOv2	decision level	Transformation matrix	Target detection	buses, cars, bicycles, motorcycles and pedestrians	PASCAL VOC, VOC2007, VOC2012

4 Projeto, implementação e resultados

Com o objetivo de desenvolver e avaliar a fusão de dados em níveis baixo e médio entre sensores câmera e radar no ambiente automotivo, foi escolhida a base de dados nuScenes, devido à sua popularidade no meio acadêmico e à sua rica coleção de cenas multissensoriais sincronizadas, capturadas nas cidades de Boston e Cingapura, conhecidas por seu tráfego urbano intenso e variado. Como o Detectron2 não oferece suporte nativo aos dados do nuScenes, adotamos o formato COCO (Common Objects in Context) como padrão de anotação. A conversão para esse formato foi necessária para viabilizar a utilização da base em frameworks de detecção de objetos 2D.

Para a implementação, treinamento e avaliação dos modelos, utilizamos a biblioteca de código aberto Detectron2 (WU, Y. et al., 2019). O Detectron2 fornece uma infraestrutura eficiente e modular para tarefas de visão computacional, permitindo o treinamento de diversos modelos de detecção e segmentação, como Faster R-CNN, Mask R-CNN e RetinaNet. Além disso, conta com o Detectron2 Model Zoo, um repositório de modelos pré-treinados que facilita a transferência de aprendizado, oferecendo diversas combinações de backbones, como R50-FPN, R101-C4 e R101-DC5. Para garantir a eficiência da detecção baseada em câmera, adotou-se a estratégia de transferência de aprendizado, uma vez que o treinamento com pesos aleatórios na base nuScenes exigiria mais tempo de processamento e não atingiria, o mesmo nível de acurácia de modelos inicializados com pesos treinados em conjuntos maiores, como COCO ou ImageNet. A Figura 10 organiza visualmente o fluxo de atividades conduzidas no desenvolvimento do trabalho, antecipando os tópicos descritos nesta seção.

Figura 10 – Fluxograma do desenvolvimento

Fonte: Produzido pelo autor

4.1 Visão Geral

A primeira proposta de fusão, em nível baixo, teve como base a adaptação do RRPN ao Detectron2, à qual foi incorporada a extensão desenvolvida neste trabalho: um gerador de âncoras customizado, definido a partir de um estudo das características geométricas das categorias presentes na base nuScenes, de forma a adequar melhor as âncoras aos tipos de objetos buscados. Para a avaliação, três modelos foram treinados: o Faster R-CNN tradicional, utilizado como referência por considerar apenas a imagem da câmera; o modelo com RRPN, no qual os dados de radar são incorporados ao processo de geração de propostas; e a versão estendida da RRPN, correspondente à proposta deste trabalho, que inclui o gerador de propostas customizado.

A segunda proposta de fusão, em nível médio, consistiu em modificações estruturais no backbone da rede, de forma a permitir a integração direta entre os sinais de radar e as imagens da câmera. A arquitetura resultante combina duas redes ResNet-50 paralelas, sendo uma responsável pelo processamento da câmera e a outra pelo radar, cujas saídas são fundidas por meio de um módulo de atenção espacial (*Spatial Attention Fusion* – SAF) (CHANG et al., 2020), aplicado antes da pirâmide de características (FPN). Para a avaliação dessa abordagem, dois modelos foram treinados: o Faster R-CNN tradicional, utilizado como referência por considerar apenas imagens da câmera, e a arquitetura proposta, que integra os dados de câmera e radar por meio da fusão SAF.

O processo de avaliação teve como objetivo quantificar os ganhos na tarefa de detecção de objetos 2D proporcionados pela fusão sensorial entre câmera e radar. Para isso, foram utilizadas as métricas oficiais do conjunto COCO. Com o intuito de investigar a robustez dos modelos em condições ambientais adversas, a base de validação foi segmentada em três subconjuntos: validação completa, cenas noturnas e cenas com chuva. Essa divisão teve como propósito avaliar o desempenho dos modelos em diferentes cenários de visibilidade.

4.1.1 Base de dados nuScenes

O nuScenes(CAESAR et al., 2020) é um banco de dados público desenvolvido pela empresa nuTonomy, lançado em 2019. Ele oferece um conjunto abrangente de dados multi-modais, contendo 1.000 cenas capturadas em Boston e Cingapura, cidades conhecidas por seu tráfego intenso e desafios complexos de direção. Cada cena tem uma duração de 20 segundos, contendo amostras sincronizadas de imagens, LIDAR e RADAR, adquiridas a uma taxa de 2 Hz. Para a coleta dos dados, foram utilizados dois veículos Renault Zoe, equipados com um conjunto idêntico de sensores, conforme ilustrado na Figura 11 e detalhado na Tabela 9. O conjunto de dados resultante inclui aproximadamente 1,4 milhão de imagens de câmera, 390 mil varreduras LIDAR e 1,4 milhão de varreduras RADAR, além de 1,4 milhão de anotações de objetos distribuídas em 23 classes.

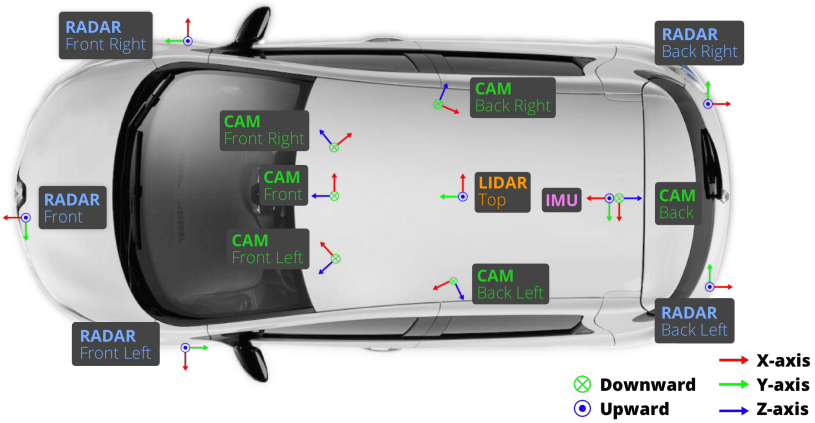

Figura 11 – Configuração dos sensores no veiculo NuScenes.

Fonte: (CAESAR et al., 2020)

Tabela 9 – Sensores utilizados pela nuScenes.

Sensor	**Detalhes**
6x Câmera	RGB, frequência de captura de 12Hz, sensor CMOS de 1/1.8", resolução de 1600 × 900, auto exposição, comprimido em JPEG
1x Lidar	Giratório, 32 feixes, frequência de captura de 20Hz, FOV horizontal de 360°, FOV vertical de -30° a 10°, alcance de ≤ 70m, precisão de ±2cm, até 1.4M pontos por segundo
5x Radar	Alcance de ≤ 250m, 77GHz, FMCW, frequência de captura de 13Hz, precisão de vel. de ±0.1km/h
GPS & IMU	GPS, IMU, AHRS. Precisão de 0.2° em orientação, 0.1° em rotação/balanço, posicionamento RTK de 20mm, taxa de atualização de 1000Hz

Os objetos são anotados quando contêm pelo menos um ponto detectado por LiDAR ou radar. Cada anotação inclui a categoria semântica do objeto, atributos como visibilidade e pose, além de *cuboids* que descreve sua posição e dimensões 3D, representado pelos parâmetros, largura, comprimento, altura e ângulo de guinada. As detecções de radar são armazenadas em um conjunto de 18 campos, que descrevem a posição, velocidade e qualidade da detecção dos objetos. Entre os principais, destacam-se as coordenadas x, y e z, que representam a posição tridimensional do ponto, enquanto v_X e v_Y indicam suas velocidades em metros por segundo. As velocidades v_{Xcomp} e v_{Ycomp} são corrigidas para compensar o movimento do veículo. O campo RCS (Radar Cross Section) quantifica a intensidade do sinal refletido, indicando a capacidade do objeto de refletir ondas de radar. Além disso, diversos outros campos fornecem informações sobre a validade da detecção e as incertezas associadas às medições.

Para utilizar os dados dos sensores em referencial comum, garantindo a correta fusão de dados, a base nuScenes disponibiliza o campo calibrated_sensor, que contém as tabelas de rotação e translação de cada sensor, além da matriz intrínseca das câmeras, obtidas a partir de um processo de calibração realizado aproximadamente duas vezes por semana ao longo dos seis meses de coleta de dados. A rotação e a translação permitem alinhar os sensores ao referencial do veículo, enquanto a matriz intrínseca possibilita a projeção da nuvem de pontos 3D nas imagens da câmera.

4.1.2 Conversão para o formato COCO

O COCO organiza as informações em um arquivo JSON estruturado em três seções principais: images, categories e annotations. A seção images armazena metadados sobre cada imagem do conjunto de dados, incluindo um identificador único (id), o nome do arquivo e suas dimensões em pixels. A seção categories define as classes dos objetos detectados, que, para esse nossos modelos serão: Pedestre, Bicicleta, Motocicleta, Carro, Caminhão e Ônibus. Já a seção annotations contém as informações de cada objeto anotado dentro das imagens, vinculando cada anotação a uma imagem específica por meio do campo image_id, correspondente ao id da imagem. Além disso, cada anotação inclui o identificador da categoria (category_id), as coordenadas da caixa delimitadora (bbox) e a área ocupada pelo objeto.

Para incorporar os dados do radar, adicionamos a seção pointcloud, que contém os campos de identificação (id), referência à imagem correspondente (image_id) e os dados de aferição (points). O campo points armazena uma lista de todos os pontos de radar associados a cada imagem, incluindo as coordenadas projetadas no plano da imagem (x e y em pixels), a distância do ponto ao radar (em metros) e as velocidades relativas (vx e vy em metros por segundo).

A conversão das detecções do radar para o plano da imagem no nuScenes é realizada utilizando o devkit do conjunto de dados, que aplica uma série de transformações baseadas nas matrizes de calibração dos sensores, garantindo o alinhamento espacial e temporal por meio do registro de data e hora de cada sensor. Inicialmente, os pontos do radar são carregados no referencial do próprio sensor e transformados para o referencial do veículo no timestamp correspondente à captura do radar. Em seguida, os pontos são convertidos para o sistema global e, posteriormente, ajustados para o referencial do veículo no instante da captura da imagem da câmera. Depois disso, a transformação para o referencial da câmera é aplicada, e a projeção dos pontos no plano da imagem ocorre por meio da matriz intrínseca da câmera. Por fim, pontos fora do campo de visão ou atrás da câmera são descartados, garantindo uma fusão precisa dos dados radar-visão.

4.1.3 Configuração do treinamento e Avaliação

O treinamento dos modelos foi realizado utilizando dados da câmera e do radar frontal provenientes do banco de dados nuScenes (CAESAR et al., 2020), previamente convertidos para o formato COCO (LIN; MAIRE et al., 2014). A Tabela 10 apresenta a configuração adotada para o treinamento dos modelos avaliados, contemplando diferentes níveis de fusão sensorial. Nela, são especificados os principais hiperparâmetros utilizados, incluindo o gerador de propostas, a arquitetura de backbone, o número máximo de iterações, a taxa de aprendizado inicial (LR Base), o otimizador, os passos de ajuste da taxa de aprendizado (Steps) e o fator de decaimento (Gamma). Também são indicados o limiar de pontuação para os testes (Score Thresh Test), o número de classes e os tipos de objetos detectados. Cada abordagem de fusão é comparada com o benchmark Faster R-CNN, identificado pelo símbolo (*), com o objetivo de avaliar os ganhos proporcionados pelas estratégias de fusão de dados e de características.

Tabela 10 – Configuração de treinamento dos modelos por nível de fusão

Configurações	**Fusão de Dados**	**Fusão de Características**
Gerador de Propostas	RPN* |RRPN |**Custom_RRPN**	RPN
Backbone	R50-FPN	R50-FPN* |**2R50-SAF-FPN**
Máx Iter	30000	20000
LR Base	0.0005	0.00025
Otimizador	-	AdamW
Steps	25000; 28000	-
Gamma	0.1	-
Score Thresh Test	0.6	0.6
Num Classes	06	06
Classes	Pedestre, Bicicleta, Motocicleta, Carro, Caminhão e Ônibus	

* Modelo de referência, Faster-RCNN (Benchmark).

Para avaliar a eficiência da fusão câmera-radar na detecção de objetos 2D, utilizamos as métricas do conjunto de dados COCO (LIN; MAIRE et al., 2014). Dentre essas métricas, destacam-se a *Precisão Média* (AP), que quantifica a capacidade do modelo de realizar predições corretas ao calcular a área sob a curva de precisão versus revocação, e a *Revocação Média* (AR), que expressa a capacidade de recuperar objetos anotados, ou seja, a fração de verdadeiros positivos em relação ao total de anotações. Ambas são calculadas sobre 10 limiares de *Interseção sobre União* (IoU), uniformemente espaçados de 0.50 a 0.95, com incremento de 0.05. Também utilizamos valores específicos de AP, como o AP_{50} e o AP_{75}, que correspondem aos limiares fixos de IoU 0.50 e 0.75, respectivamente. Essas métricas podem ser reportadas de forma global ou segmentadas por categoria. Conforme a convenção do COCO, não se distingue AP de mAP (assim como AR de mAR), presumindo-se que o contexto torne essa equivalência clara. Para uma análise em diferentes escalas, também reportamos os valores de AP segmentados pelo tamanho dos objetos, conforme a definição do COCO: pequenos (AP_s, área $< 32^2$), médios (AP_m, $32^2 <$ área $< 96^2$) e grandes (AP_l, área $> 96^2$).

Com o objetivo de avaliar o desempenho dos modelos em diferentes condições de condução, o conjunto de validação foi segmentado em três grupos: dia, noite e chuva. A separação entre cenas diurnas e noturnas baseou-se no horário de captura, enquanto as cenas de chuva foram identificadas pela chave *rain* na descrição de cena. A Tabela 10 apresenta a distribuição das instâncias entre essas categorias. Devido à alta similaridade entre o conjunto completo e o grupo diurno, optou-se por não utilizar este último na avaliação.

Tabela 11 – Distribuição das instâncias entre todas as 6 categorias

Categoria	**Total (N)**	**Dia**	**Noite**	**Chuva**
Pessoa (N)	8,635	8,598	37	382
Bicicleta (N)	556	552	4	34
Motocicleta (N)	922	753	169	70
Carro (N)	22,236	20,722	1,514	5,166
Ônibus (N)	1,162	1,162	0	169
Caminhão (N)	4,640	4,558	82	1,281
Total (N)	38,151	36,345	1,806	7,102
Total (%)	100	95.3	4.7	18.6

4.2 Método 1: Fusão em nível de dados baseada em RRPN

Esta seção explora o método de fusão sensorial de baixo nível que utiliza a detecção por radar para gerar regiões de interesse (ROIs) no sistema de coordenadas da câmera. Para isso, foi adotado o algoritmo RRPN (*Radar Region Proposal Network*) (NABATI; QI, 2019), que gera caixas delimitadoras (âncoras) sobre a imagem a partir das coordenadas e distâncias fornecidas por cada ponto de detecção do radar.

Nabati e Qi (2019) propôs um método para ajustar fator de escala das âncoras com base na distância dos objetos detectados. Esse mecanismo parte do princípio de que objetos mais distantes ocupam áreas menores na imagem e, portanto, devem ter âncoras proporcionais à sua projeção. Cada ponto de radar é mapeado para o plano da imagem como um vetor de três componentes: coordenadas projetadas e distância. O fator de escala aplicado à âncora de cada detecção é calculado segundo a seguinte equação:

$$S_i = \alpha \frac{1}{d_i} + \beta \tag{4.1}$$

Em que S_i representa o fator de escala da âncora da detecção i, d_i é a distância ao objeto, e α e β são parâmetros definidos para ajustar esse escalonamento. Os parâmetros α e β foram determinados usando uma pesquisa de grade dentro de um intervalo de valores que maximiza a Interseção sobre União (IoU) entre as caixas delimitadoras geradas e as caixas delimitadoras da verdade básica do conjunto de treinamento (NABATI; QI, 2019).

Além da compensação baseada na distância, o algoritmo também permite configurar parâmetros como os formatos das âncoras, definidos por *aspect ratios*, as posições relativas ao ponto de detecção (como centro, cima, esquerda e direita), e um fator multiplicador aplicado a cada âncora gerada. Essa flexibilidade possibilita a criação de um gerador de âncoras customizado, conforme será apresentado nas seções seguintes.

4.2.1 Gerador de Âncora RRPN

As âncoras são geradas com base nos sinais de radar, para cada leitura de radar no plano da imagem, são adotados três aspect ratios, que representam a razão entre a altura e a largura das caixas delimitadoras: 0,5, 1 e 2. Esses valores correspondem, respectivamente, a retângulos horizontais, quadrados e retângulos verticais, permitindo uma melhor adaptação das âncoras às diferentes formas dos objetos presentes na cena.

Como os pontos de radar raramente coincidem exatamente com o centro dos objetos de interesse, as âncoras não são geradas apenas na posição central do ponto de detecção, mas também deslocadas para as posições superior, esquerda e direita. Dessa forma, cada ponto de detecção do radar contribui com quatro pontos distintos de referência para a geração de âncoras, conforme ilustrado na Figura 12.

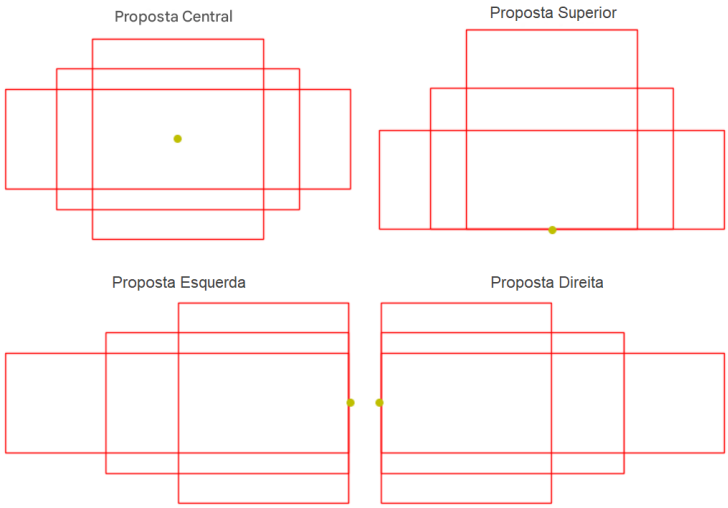

Figura 12 – Âncoras geradas pelo RRPN.

Fonte: Produzido pelo autor

Além das variações de posicionamento, são aplicados fatores de escala multiplicativos de 1×, 2× e 4×, ajustando o tamanho das caixas delimitadoras para abranger objetos de diferentes dimensões. Assim, considerando as três proporções de *aspect ratio* e as quatro posições de deslocamento, a introdução dos fatores de escala resulta em um total de 36 caixas delimitadoras por ponto de detecção do radar, como ilustrado na Figura 13.

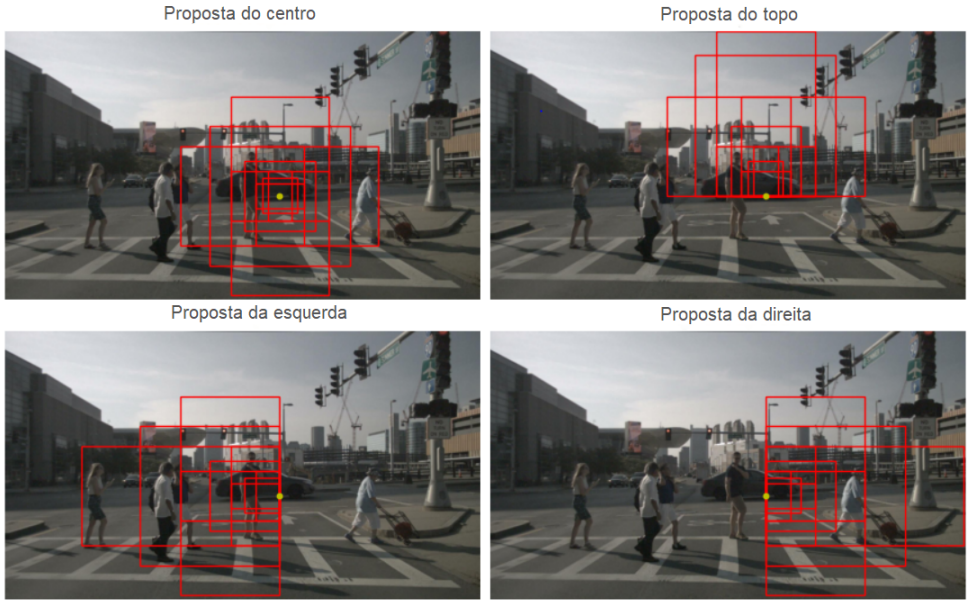

Figura 13 – Bounding boxes geradas por ponto de radar

Fonte: Produzido pelo autor com dados de nuScenes

4.2.2 Gerador de âncora customizado

Para obter âncoras mais representativas para os tipos de objetos que se pretende identificar, sendo eles: pedestres, bicicletas, motocicletas, carros, ônibus e caminhões. Investigou-se as proporções das caixas delimitadoras da verdade básica no conjunto de treinamento NuScenes(CAESAR et al., 2020). A relação entre a altura e a largura das caixas delimitadoras de cada objeto de interesse foi extraída e agrupada em três categorias distintas: Pequeno para pedestres, bicicletas e motocicletas; Médio para carros; e Grande para ônibus e caminhões. Cada categoria foi plotada em um histograma para analisar as proporções mais frequentes por tamanho do objeto conforme mostrado na Figura 14.

A análise do histograma da Figura 14, mostra que objetos menores tendem a ter uma caixa delimitadora retangular vertical, com uma proporção de cerca de 1,8. As distribuições das barras nos histogramas para objetos médios e grandes são mais semelhantes, com objetos médios tendo uma proporção de aspecto mais padronizada, com numerosos picos em torno de 0,4, 0,6 e 0,8. Da mesma forma, objetos grandes têm uma frequência mais baixa e um desvio padrão mais alto. Dadas essas características de cada objeto, foram escolhidos os seguintes multiplicadores e proporções para gerar as âncoras do modelo proposto.

Tabela 12 – Aspect Ratios for Different Multiplied Factors

Fator multiplicador	**Aspect Ratio [H/W]**
1X	1.8; 0.6
2X	1.8; 0.6
3X	0.8; 0.6; 0.4
4X	0.8; 0.6; 0.4

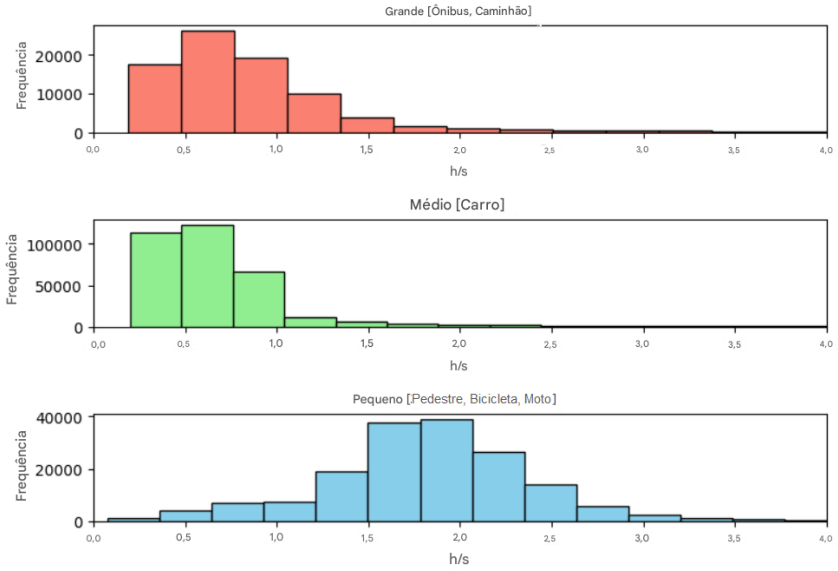

Figura 14 – Histograma do aspect ratio por tamanho de objeto

Fonte: Produzido pelo autor

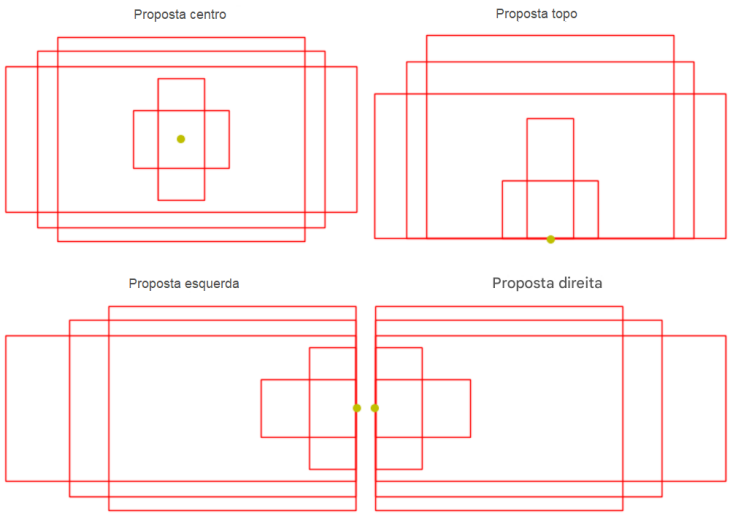

Figura 15 – Âncoras geradas pelo RRPN customizado.

Fonte: Produzido pelo autor

A Figura 15 ilustra as âncoras geradas com os fatores multiplicadores 1× e 3×, posicionadas nos pontos central, superior, esquerdo e direito em relação à detecção do radar. O processo proposto de geração de âncoras resultou em um total de 40 caixas delimitadoras por ponto de detecção. As Figuras 32 e 33, apresentadas em anexo, comparam o gerador de âncoras RRPN com a versão customizada, evidenciando as propostas de região produzidas para um mesmo ponto de detecção.

4.2.3 Resultados

Nesta seção, são apresentados os resultados obtidos no treinamento de três modelos distintos. O primeiro modelo, Faster R-CNN (REN et al., 2015), utiliza exclusivamente imagens da câmera frontal e emprega o gerador de propostas RPN. Esse modelo serve como referência para comparação, uma vez que a principal diferença entre as abordagens analisadas está no mecanismo de geração de propostas. Os outros dois modelos incorporam a fusão de dados entre câmera e radar. O segundo modelo utiliza o gerador de propostas RRPN, cuja abordagem foi detalhada na Seção 4.2.1. Por fim, o terceiro modelo, denominado *Custom_RRPN*, adota uma estratégia personalizada de geração de âncoras, conforme descrito na Seção 4.2.2.

Os resultados experimentais são apresentados nas Tabelas 13 e 14. Além disso, as Tabelas 15 e 16 mostram os desempenhos específicos dos modelos em condições desafiadoras, como cenas noturnas e ambientes chuvosos, respectivamente.

Tabela 13 – Precisão média por categoria na base completa

Categoria	Faster R-CNN	RRPN	Custom RRPN
Pessoa (AP)	15.480	7.734	6.452
Bicicleta (AP)	7.434	5.604	4.978
Motocicleta (AP)	8.898	5.515	6.869
Carro (AP)	36.639	28.566	28.355
Ônibus (AP)	34.139	29.622	34.916
Caminhão (AP)	16.637	13.332	16.555

Tabela 14 – Resultados na base de validação completa

Métrica	Faster R-CNN	RRPN	Custom RRPN
AP	19.87	15.22	16.35
AR	25.4	20.0	21.0
AP50	38.79	30.02	31.05
APs	1.60	0.36	0.21
APm	11.66	6.23	6.18
AP1	30.9	25.78	27.66

Os resultados nas Tabelas 13 a 16 destacam uma melhoria nas métricas entre o Custom_RRPN em comparação com o RRPN, porém não muito significativa. Comparado ao Faster R-CNN (REN et al., 2015), todas as métricas foram inferiores neste treinamento. A técnica de fusão de dados, que utiliza radar para gerar regiões de interesse propostas para a câmera, degrada o desempenho do detector de objetos que utiliza apenas a câmera.

Tabela 15 – Resultados na base de validação noturna

Métrica	Faster R-CNN	RRPN	Custom RRPN
AP	12.94	9.07	9.58
AR	16.0	11.9	11.6
AP50	24.39	18.96	19.26
APs	6.13	0.03	0.17
APm	9.75	4.42	5.21
AP1	19.01	15.10	16.26

Tabela 16 – Resultados na base de validação em condição de chuva

Métrica	Faster R-CNN	RRPN	Custom RRPN
AP	13.67	10.56	11.78
AR	17.6	14.2	15.5
AP50	28.22	21.48	23.89
APs	10.46	7.32	7.56
APm	10.27	6.39	7.54
AP1	20.57	15.39	18.30

Neste formato de fusão, o radar é utilizado como sensor principal para a detecção. Dessa forma, objetos não detectados pelo radar não são procurados nem analisados pela rede neural. Para validar essa hipótese, os objetos anotados nos conjuntos de validação e teste foram carregados. Para cada objeto, foi verificado se havia pontos de radar associados e quantificada a quantidade de pontos relacionados. Manteve-se a mesma divisão de categorias na análise da proporção por tamanho de objeto: a Figura 16 apresenta os resultados para objetos pequenos, a Figura 17 para objetos médios e a Figura 18 para objetos grandes.

Os histogramas apresentados nas Figuras 16 a 18 indicam que objetos pequenos possuem maior probabilidade de não serem detectados pelo radar. Quando detectados, raramente apresentam mais de um ponto vinculado. No caso específico da detecção de pedestres, aproximadamente 20% dessa categoria possui ao menos um ponto de radar associado, o que justifica, os baixos índices de detecção observados na Tabela 13. A análise das demais categorias de objetos na cena segue um padrão semelhante, corroborando os resultados do experimento apresentados nas Tabelas 13 a 16. Conforme evidenciado, caminhões e ônibus foram melhor detectados pelo radar e apresentaram maior número de pontos associados, resultando em valores de APl mais próximos aos obtidos com o modelo Faster R-CNN (REN et al., 2015).

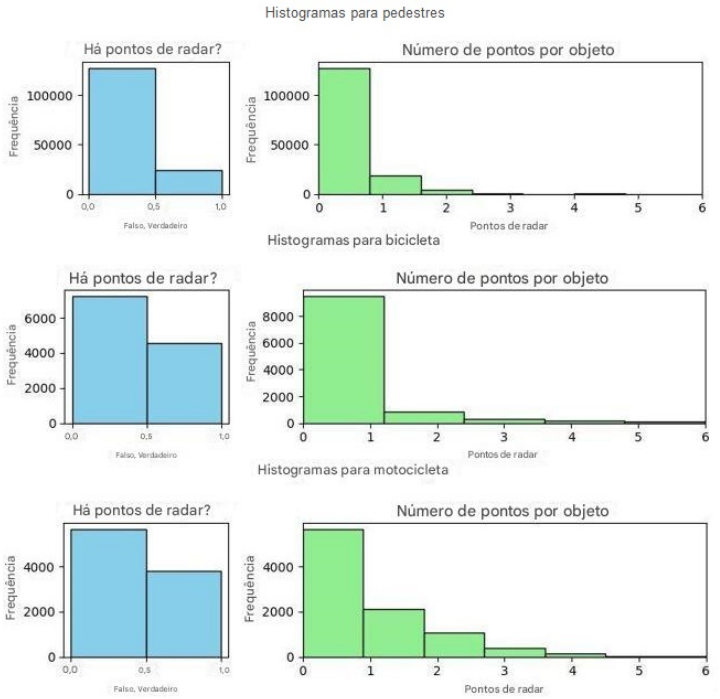

Figura 16 – Histograma de detecções de radar para pequenos objetos no banco de dados NuScenes.

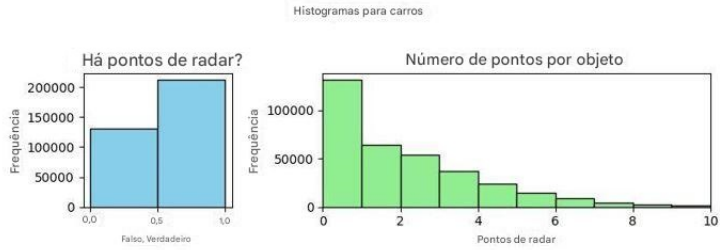

Figura 17 – Histograma de detecções de radar para objetos médios no banco de dados NuScenes.

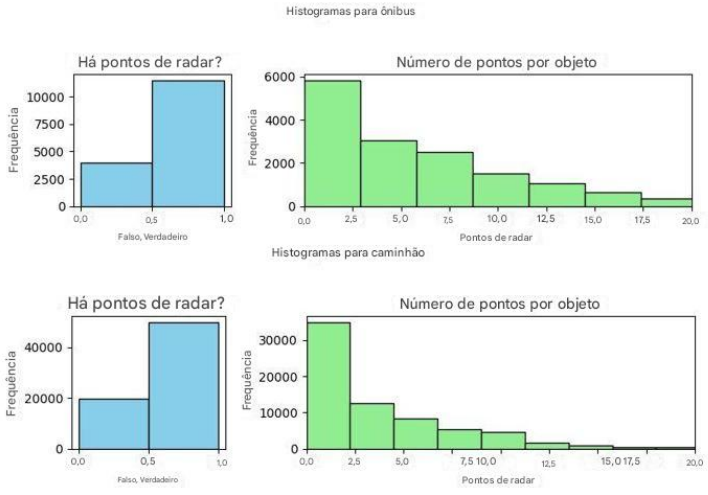

Figura 18 – Histograma de detecções de radar para objetos grandes no banco de dados NuScenes.

O algoritmo RRPN (NABATI; QI, 2019) pode aumentar potencialmente a velocidade do Fast R-CNN (GIRSHICK, 2015), aproximando-o do desempenho em tempo real exigido por aplicações em veículos autônomos. No entanto, com o lançamento do Faster R-CNN(REN et al., 2015), que integra de forma mais eficiente a geração de propostas por meio da *Region Proposal Network* (RPN), torna-se impraticável utilizar exclusivamente a detecção por radar para gerar regiões de interesse. Para mitigar a ausência de detecção em áreas sem retorno do radar, é essencial também realizar a extração de características da imagem. Dessa forma, os modelos de detecção baseados em redes neurais convolucionais (CNNs), quando implementados com esquemas de fusão de nível médio (ou fusão de características), têm se mostrado mais eficazes ao integrar informações de diferentes fontes e melhorar a robustez da detecção.

4.3 Método 2: Fusão em nível de característica baseado em SAF

Esta seção explora o método de fusão em nível de característica, com base no trabalho de Chang et al. (2020). Nesse método, a fusão por meio do módulo SAF (*Spatial Attention Fusion*) tem como principal objetivo aprimorar a detecção de pequenos objetos e de objetos desfocados, ao enfatizar regiões relevantes e reforçar a confiabilidade das detecções oriundas da câmera. Além disso, essa abordagem assegura que as áreas sem retorno do radar sejam preservadas após a fusão, permitindo que continuem sendo consideradas nas etapas subsequentes do processo de detecção. O detector de objetos adotado foi o Faster R-CNN, amplamente reconhecido como referência na área de detecção de objetos. Para aprimorar o reconhecimento em múltiplas escalas, integrou-se à arquitetura o módulo *Feature Pyramid Networks* (FPN).

4.3.1 Fluxo dos dados do radar no Detectron2

Como o Detectron2 (WU, Y. et al., 2019) é projetado exclusivamente para trabalhar com imagens, as leituras do radar precisam ser processadas e enviadas até a etapa de extração de recursos para que possam ser fundidas com as características da câmera. Após a extração dos dados da base nuScenes e a geração do arquivo JSON no formato COCO, que inclui informações do radar (coordenadas, distância e velocidades) para cada imagem, conforme detalhado na Seção 4.1.2, a função ***load_coco_json*** é utilizada para carregar esses dados. Essa função gera um dicionário contendo os metadados de cada imagem do banco (como nome do arquivo, ID e dimensões), juntamente com as anotações correspondentes (caixas delimitadoras e categorias dos objetos). Para acomodar as informações do radar, essa função foi modificada de modo a incluir os dados no formato do dicionário, conforme exemplificado em A.1.

As classes ***DatasetMapper*** e ***DefaultPredictor*** também são modificadas. Durante a execução do treinamento ou predição, essas classes são responsáveis por buscar a imagem a partir do caminho armazenado no campo "file_name", transformá-la em tensor e carregá-la no dispositivo configurado (GPU ou CPU). Além de carregar a imagem da câmera, como fazem as classes originais, a versão customizada também gera uma imagem RGB a partir dos dados do radar. Nessa conversão, as coordenadas são usadas para posicionar os pontos na imagem, enquanto os valores dos canais representam a distância, a velocidade relativa em X e a velocidade relativa em Y. Como saída, a classe retorna um dicionário contendo as imagens da câmera e do radar devidamente carregadas e transformadas em tensores.

As informações da câmera e do radar são fundidas após a extração de características de cada sensor pelo ResNet-50, e as características combinadas são então enviadas para o FPN, seguindo o fluxo normal do modelo. Toda essa fusão em nível de características ocorre na camada de backbone do Detectron2, exigindo uma adaptação na classe ***GeneralizedRCNN***, que implementa um modelo genérico baseado em R-CNN dentro da arquitetura do Detectron2. Essa classe gerencia a extração de características, a geração de propostas e a predição final em modelos como o Faster R-CNN.

Os tensores da câmera e do radar são passados para o backbone utilizando uma operação de concatenação. Para isso, o tensor de imagem do radar é redimensionado para o tamanho da imagem da câmera, já que esse é um parâmetro configurável no Detectron2. Outras operações de pré-processamento, como normalização, já aplicadas à imagem da câmera em ***GeneralizedRCNN***, também são realizadas na imagem do radar, assegurando um tratamento consistente entre os sensores. Por fim, a operação de concatenação é revertida no backbone, os tensores são enviados para o extrator de caracteristicas, e uma operação de fusão pode ser aplicada para combinar os mapas de características dos sensores.

4.3.2 Gerando imagem da nuvem de pontos do radar

As nuvens de pontos e as detecções do radar não podem ser diretamente utilizadas nos extratores de características, já que esses métodos operam em imagens. Para contornar essa situação, os dados do radar são processados e transformados em uma imagem com as mesmas proporções da imagem da câmera, mantendo a equivalência espacial entra elas. Cada canal da imagem de radar, responsável pela coloração R, G e B, é associado a um valor físico detectado. A abordagem escolhida, baseada em Chang et al. (2020), escolhemos a distância para preencher o canal vermelho, a velocidade radial para o canal verde e a velocidade transversal para o canal azul. Cada medição é representada na Figura 19, com os valores mapeados por coloração para melhor visualização. Além disso, seguimos a abordagem do artigo para representar a detecção de radar como um círculo na imagem. No estudo, foi avaliado diferentes raios entre 1 e 11 e concluíram que o raio 7 proporcionou o melhor desempenho no ResNet-50.

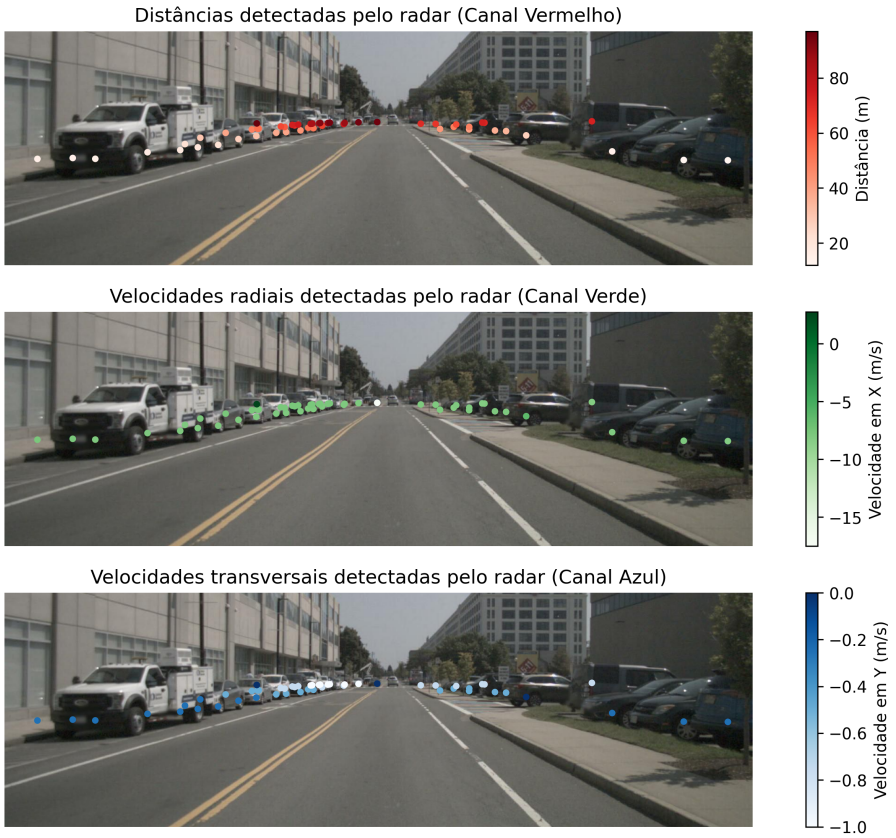

Figura 19 – Detecções de distância, velocidade radial e velocidade transversal obtidas pelo radar

Fonte: Produzido pelo autor com dados de nuScenes

Figura 20 – Imagem câmera e imagem gerada para o radar

Fonte: Produzido pelo autor com dados de nuScenes

Para obter uma regra de representação dessas medições na imagem, cada detecção é convertida em intensidade de pixel, variando de 0 a 255. A faixa de valores escolhida para a distância foi de 0 a 250 metros, e para as velocidades relativas, de -33 a 33 m/s (-118 a 118 km/h). A faixa de valores de cada medição é normalizada para que os valores em pixel fiquem entre 127 e 255, conforme a Equação 4.2. As imagens da mesma cena, obtidas pela câmera e geradas para o radar, são apresentadas na Figura 20.

$$Pixel = \left(\frac{\text{Detecção} - \text{MIN}}{\text{MAX} - \text{MIN}} \times 128 + 127 \right) \tag{4.2}$$

4.3.3 Arquitetura

A arquitetura detalhada do modelo de referência é apresentada na Figura 21. Nela, os rótulos em azul destacam as classes do Detectron2 que implementam cada módulo do Faster R-CNN com ResNet e FPN (Benchmark). A principal classe, ***GeneralizedRCNN***, estrutura o modelo em três blocos: o *backbone*, responsável pela extração de características da imagem; o *RPN*, encarregado de gerar propostas de região que podem conter objetos; e o *ROI Heads*, responsável por processar as regiões propostas, e ajustar as caixas delimitadoras.

O *backbone* utiliza a ResNet-50 para extrair características da imagem de entrada. Essas características são refinadas pela *Feature Pyramid Network (FPN)*, que combina informações de diferentes níveis da rede para gerar mapas de características em múltiplas escalas. As saídas da FPN são representadas pelos níveis $P2$, $P3$, $P4$, $P5$ e $P6$, correspondendo a resoluções reduzidas por fatores de 4, 8, 16, 32 e 64 em relação à imagem original, respectivamente.

A fusão de características entre os dados da câmera e do radar ocorre no bloco *backbone*, como falado na seção anterior, mantendo os demais blocos do modelo de referência sem alterações. O *backbone* proposto utiliza duas redes ResNet-50, uma para extrair características da imagem da câmera e outra para a imagem do radar. Essas características passam pelo bloco *Spatial Attention Fusion (SAF)* para serem fundidas antes de seguirem para o refinamento no FPN, como mostrado na Figura 22.

O bloco de fusão SAF utiliza uma matriz de atenção espacial gerada a partir das características extraídas do radar, que são aplicadas aos mapas de características da visão. Esse mecanismo realça regiões onde o radar fornece informações relevantes e atenua a influência de áreas menos confiáveis. Para isso, as características do radar passam por três camadas convolucionais com diferentes tamanhos de kernel (1×1, 3×3 e 5×5), cada uma com padding correspondente (0, 1 e 2, respectivamente), a fim de preservar as proporções espaciais das features originais. As saídas dessas convoluções são somadas, resultando em uma matriz de atenção espacial com as mesmas dimensões do mapa de características da câmera, como ilustrado na Figura 23.

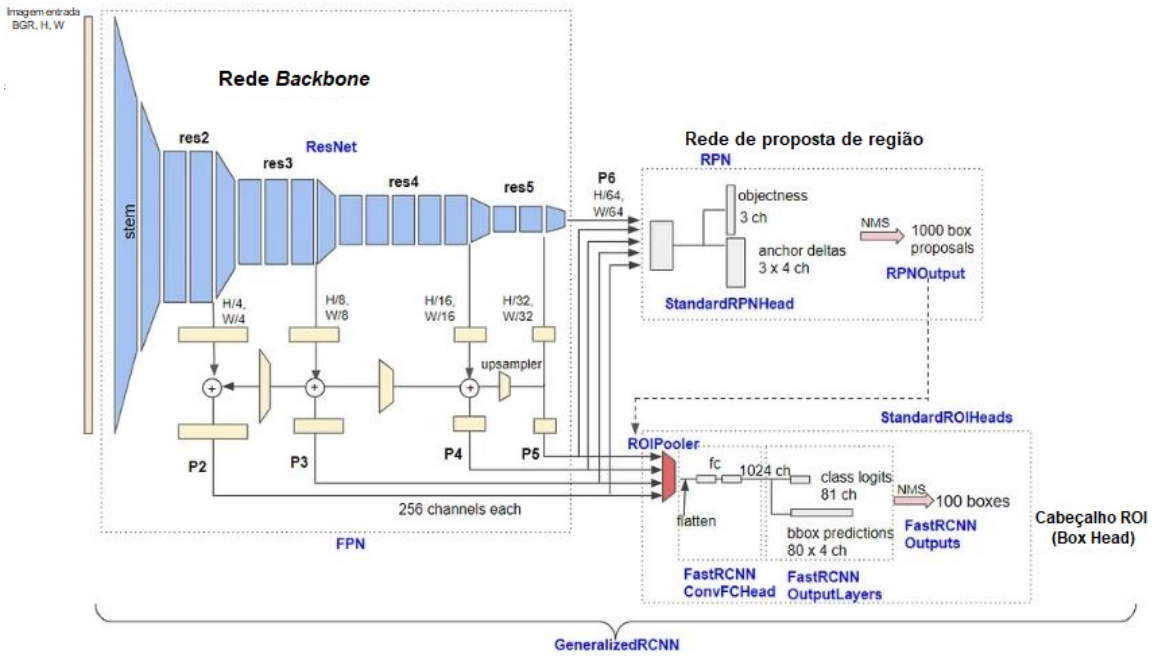

Figura 21 – Arquitetura detalhada do Faster RCNN com FPN e Resnet. Rótulos azuis representam nomes de classes no Detectron2

Fonte: (HONDA, 2020)

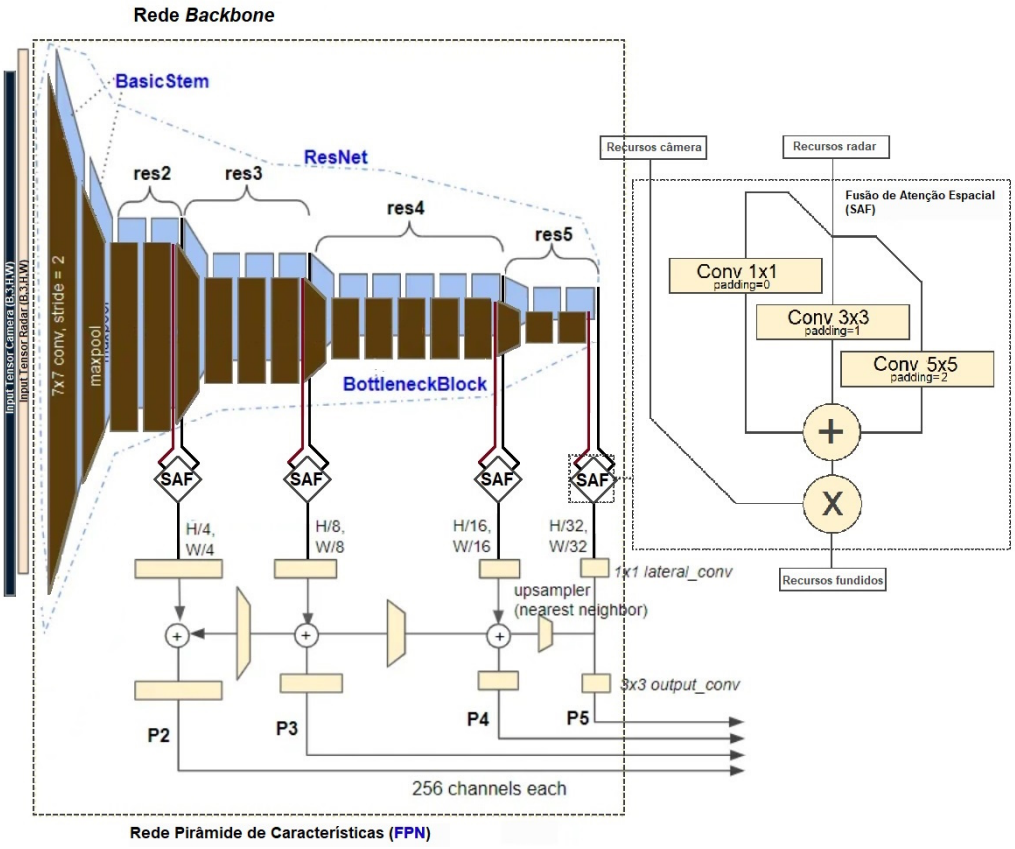

Figura 22 – Arquitetura do backbone (2R50-SAF-FPN)

Fonte: Modificado de (HONDA, 2020)

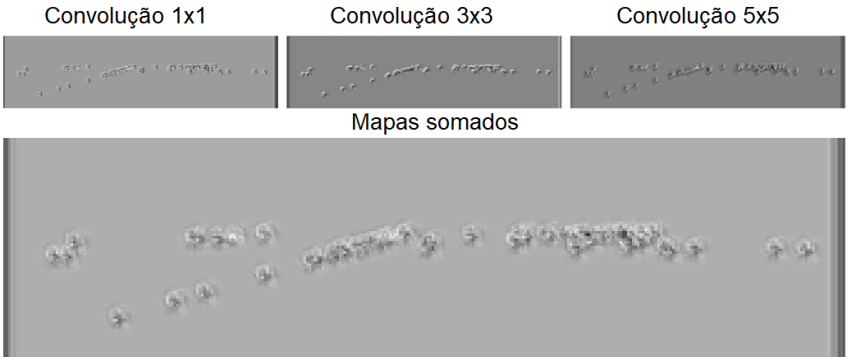

Figura 23 – Operações aplicadas no mapa de característica do radar

Fonte: Produzido pelo autor com dados de nuScenes

Para concluir a fusão SAF, a matriz de atenção espacial gerada a partir das características do radar é multiplicada pelos mapas extraídos da visão, modulando a informação visual com base na percepção complementar do radar. Esse processo realça regiões com retorno do radar, como objetos em movimento ou altamente refletivos, sem suprimir as áreas da imagem onde não há detecção. A Figura 24 ilustra esse efeito, comparando os mapas de características da câmera (linha superior), do radar (linha intermediária) e da fusão SAF (linha inferior), nas saídas res2, res3 e res4 do extrator de características. Observa-se que as regiões com retorno do radar apresentam maior ativação nos mapas fundidos, especialmente nas saídas de maior resolução, como res2 e res3, ao mesmo tempo em que a estrutura visual é preservada nas regiões sem detecção.

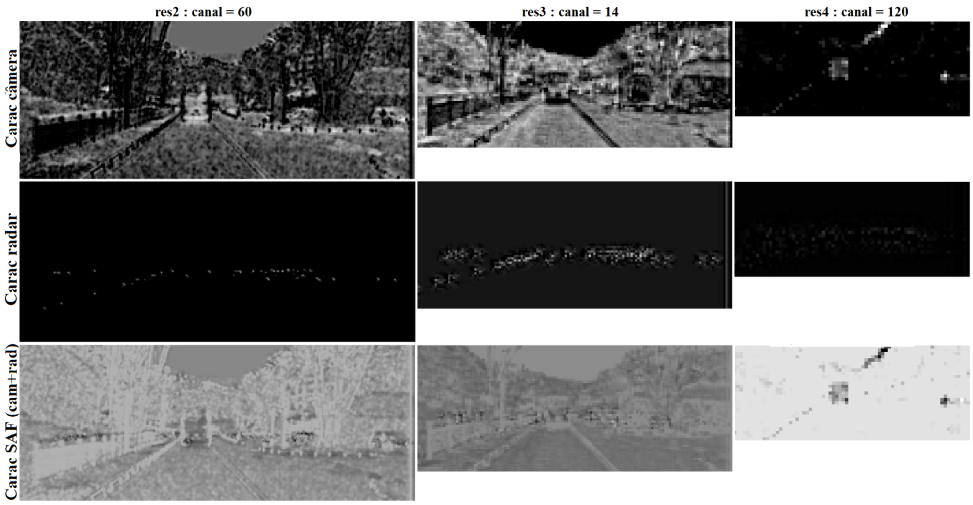

Figura 24 – Mapas de características extraídos da câmera, do radar e da fusão SAF

Fonte: Produzido pelo autor com dados de nuScenes

4.3.4 Resultados

Nesta seção, apresentamos os resultados obtidos para os dois modelos treinados. O primeiro modelo, o Faster R-CNN (REN et al., 2015), utiliza exclusivamente imagens da câmera frontal do conjunto de dados. Ele emprega a arquitetura ResNet-50 com FPN (R50-FPN) como backbone e serve como base de comparação, já que a principal diferença entre os modelos está na estrutura do backbone. No segundo modelo, realiza-se a fusão de características diretamente no mapa de características, combinando informações provenientes da câmera e do radar frontal. Para isso, são utilizadas duas redes ResNet-50, com a fusão SAF aplicada antes da FPN (2R50-SAF-FPN), conforme descrito na Seção 4.3.3.

A Figura 25 ilustra a efetividade do modelo proposto. Na primeira linha, apresenta-se a imagem de referência com os pontos de radar sobrepostos, as detecções do modelo Faster R-CNN representadas em verde e as anotações manuais (Ground Truth) em vermelho. Observa-se um veículo preto parcialmente visível, cuja frente e traseira estão encobertas por uma árvore e por uma placa de sinalização. Apesar da oclusão, diversos pontos de radar registraram a presença do veículo, o que aumentou a confiabilidade da detecção. Como resultado, a instância ultrapassou o limiar de pontuação adotado no teste (Score Threshold Test, especificado na tabela 10), sendo corretamente detectada pelo modelo Faster SAF-CNN, como ilustrado na imagem da segunda linha.

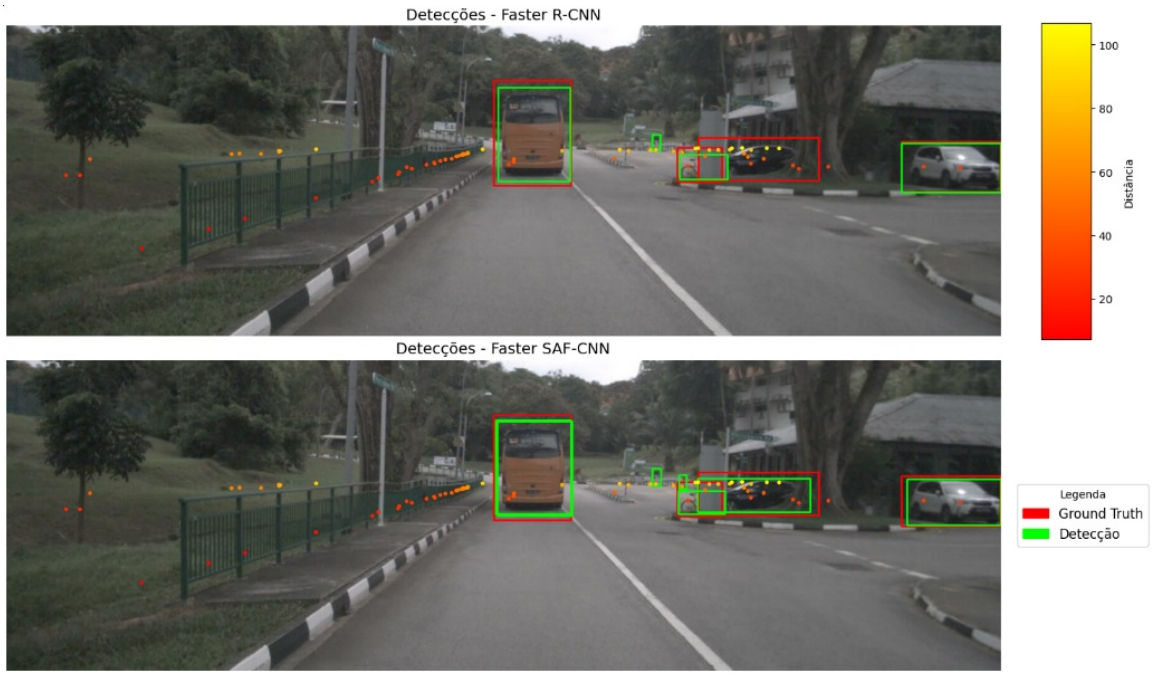

Figura 25 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

Os resultados experimentais completos são apresentados nas Tabelas 17 e 18. Além disso, as Tabelas 20 e 21 detalham o desempenho dos modelos em condições adversas, como cenas noturnas e chuvosas, respectivamente.

Tabela 17 – Precisão média por categoria na base de validação completa

Categoria	Faster R-CNN	Faster SAF-CNN
Pessoa (AP)	23.230	23.440
Bicicleta (AP)	13.448	13.301
Motocicleta (AP)	13.159	14.008
Carro (AP)	43.049	44.274
Ônibus (AP)	33.361	32.266
Caminhão (AP)	18.874	21.963

Tabela 18 – Resultados na base de validação completa

Métrica	Faster R-CNN	Faster SAF-CNN
AP	24.19	24.87
AR	33.7	34.5
AP50	47.69	48.65
AP75	21.37	23.01
APs	3.06	4.42
APm	18.53	19.02
AP1	33.56	34.30

Na avaliação sobre a base de validação completa, observa-se um incremento na precisão média de 3,09%, 1,09% e 0,85% nas categorias caminhão, carro e motocicleta (Tabela 17), respectivamente. Também foram registrados acréscimos de 0,80% em AR, 0,96% em AP50 e 1,36% em APs, indicando maior consistência na detecção de objetos dentro dos limiares estabelecidos e melhor desempenho em alvos de menor dimensão. O valor de AP75 apresentou uma elevação de 1,64%, evidenciando maior precisão na localização das caixas delimitadoras. A Tabela 19 ilustra a relevância desse resultado, mostrando que o modelo proposto identificou 383 objetos adicionais na cena de trânsito em relação ao Faster R-CNN e reduziu 703 detecções falsas, o que reforça o impacto prático dessa melhoria.

Tabela 19 – Comparação de Detecções (IoU 75%)

Modelo	**Verdadeiros Positivos**	**Falsos Positivos**	**Falsos Negativos**
Faster SAF-CNN	14,719	25,273	23,432
Faster R-CNN	14,336	25,976	23,815
Diferença	383	-703	-383

A avaliação em condições adversas, conforme apresentado nas Tabelas 20 e 21, revela uma leve superioridade do modelo proposto nas métricas analisadas. Destaca-se, em particular, um ganho de 1,35% em AP75 na base sob condição de chuva, indicando maior precisão na localização das caixas delimitadoras. No caso da base de validação em condição noturna, que representa apenas 4,7% das categorias da base completa, o modelo, embora também utilize os dados provenientes do radar, cuja resposta não é afetada pelas condições de iluminação, não apresentou um ganho tão expressivo. Esse resultado pode estar associado à reduzida quantidade de objetos detectáveis e ao número limitado de cenas noturnas, conforme mostrado na Tabela 11.

Tabela 20 – Resultados na base de validação noturna

Métrica	Faster R-CNN	Faster SAF-CNN
AP	13.22	13.40
AR	18.20	18.30
AP50	28.04	28.33
AP75	11.20	11.44
APs	11.24	11.17
APm	10.20	10.34
AP1	17.68	17.93

Tabela 21 – Resultados na base de validação em condição de chuva

Métrica	Faster R-CNN	Faster SAF-CNN
AP	18.67	18.72
AR	27.70	27.40
AP50	35.89	35.95
AP75	16.91	18.26
APs	3.85	3.60
APm	15.92	15.52
AP1	23.43	23.37

A Tabela 22 compara o tempo médio de execução entre o nosso método proposto (SAF) e o Faster R-CNN (FAS). O SAF apresentou aproximadamente três vezes o tempo do FAS, sendo 13,37% desse aumento associado ao gerador de imagens de radar, implementado em Python, e passível de otimização com linguagens mais eficientes como C++. As operações de fusão não impactaram significativamente o tempo, já que o backbone apresentou desempenho similar ao do FAS. A maior parte do tempo adicional (71,38%) foi consumida pelo gerador de propostas, resultado da fusão e modificação do mapa de características que alimenta essa etapa. Além disso, valores elevados nos mapas de características parecem influenciar o tempo gasto, sugerindo a necessidade de estudar a relação entre a ativação das regiões de interesse do radar e o desempenho do gerador de propostas, visando otimizar o equilíbrio entre precisão e velocidade.

Tabela 22 – Comparação de tempo médio entre Faster R-CNN e Faster SAF-CNN

Etapa	**FAS (ms)**	**SAF (ms)**	**SAF / FAS**	**Diferença relativa (%)**
Create_radar_img	2.97	499.16	168.0x	13.37%
Preprocessing	22.32	21.26	0.95x	-0.05%
Backbone (Fusion)	19.64	23.15	1.18x	0.14%
Proposal_Generator	1257.41	3583.87	2.85x	71.38%
Roi_Heads	44.44	47.94	1.08x	1.16%
Tempo médio	1434.12	4268.69	2.98x	100%

Os resultados obtidos evidenciam que a fusão de características por meio do módulo SAF proporciona uma melhora no desempenho da detecção de objetos, com destaque para o aumento da precisão na localização das caixas delimitadoras, sobretudo em situações envolvendo oclusões e na detecção de pequenos objetos. Essa tendência é reforçada por resultados qualitativos obtidos em diferentes condições. Em ambientes com neblina, nos quais a visibilidade da câmera é severamente reduzida, o modelo com fusão foi capaz de detectar corretamente a frente de um caminhão muito próximo e dois veículos mais distantes (Figura 26), bem como identificar um carro e uma van estacionados à frente do veículo (Figura 27).

Figura 26 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

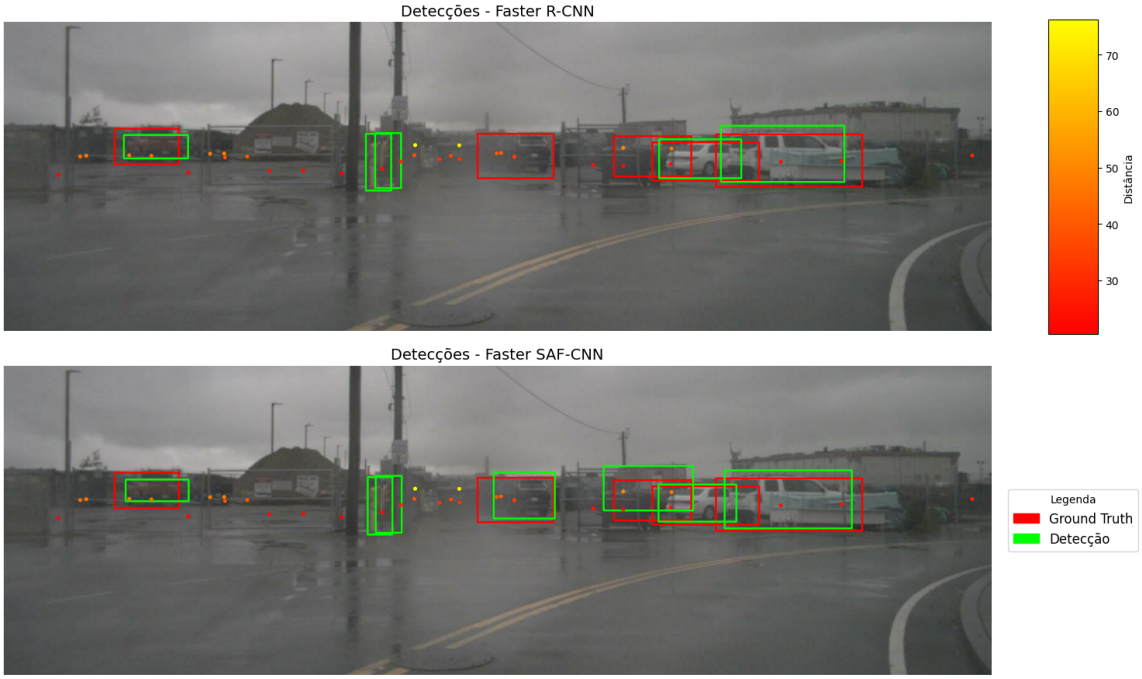

Figura 27 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

A Figura 28 destaca a capacidade do modelo em identificar um carro e um pedestre a longa distância, evidenciando sua eficácia na detecção de pequenos objetos. Na Figura 29, observa-se a correta identificação de dois caminhões muito próximos, sendo o segundo parcialmente ocultado pelo primeiro, o que demonstra a robustez do modelo frente a oclusões.

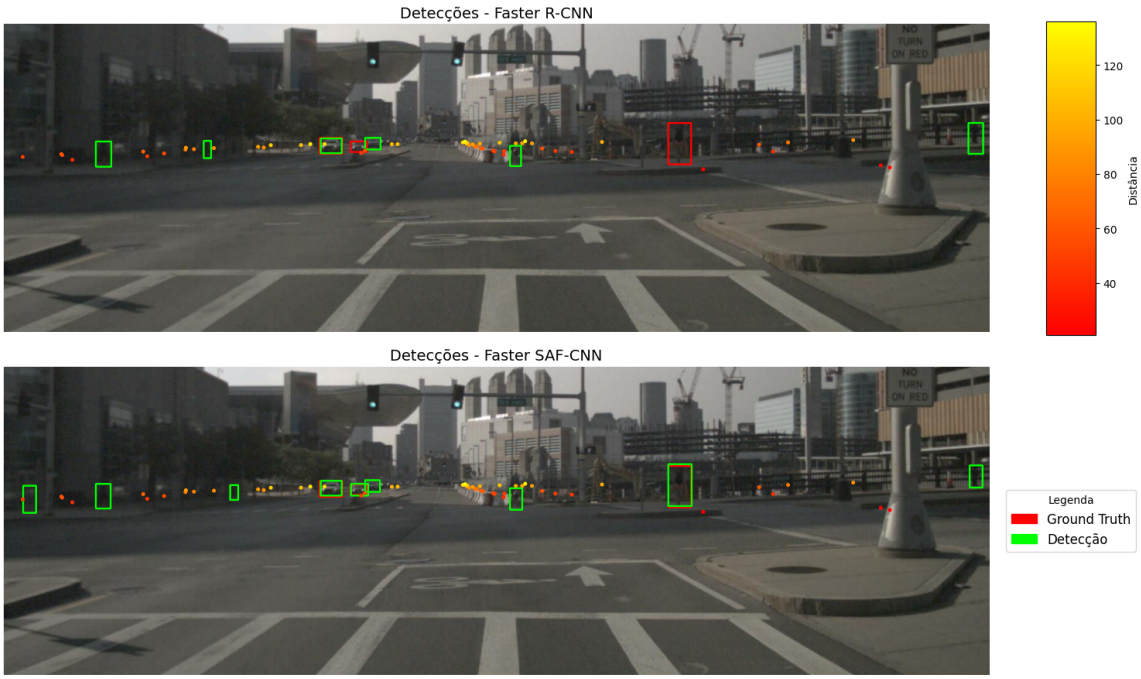

Figura 28 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

Figura 29 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

A Figura 30 apresenta um caminhão de grande porte a média distância e um pedestre à frente do veículo, enquanto a Figura 31 evidencia a presença de um caminhão de pequeno porte a curta distância, posicionado à direita. Em ambos os casos, o modelo multissensorial realiza detecções críticas do ponto de vista da segurança veicular, uma vez que os objetos identificados representam potenciais riscos de colisão.

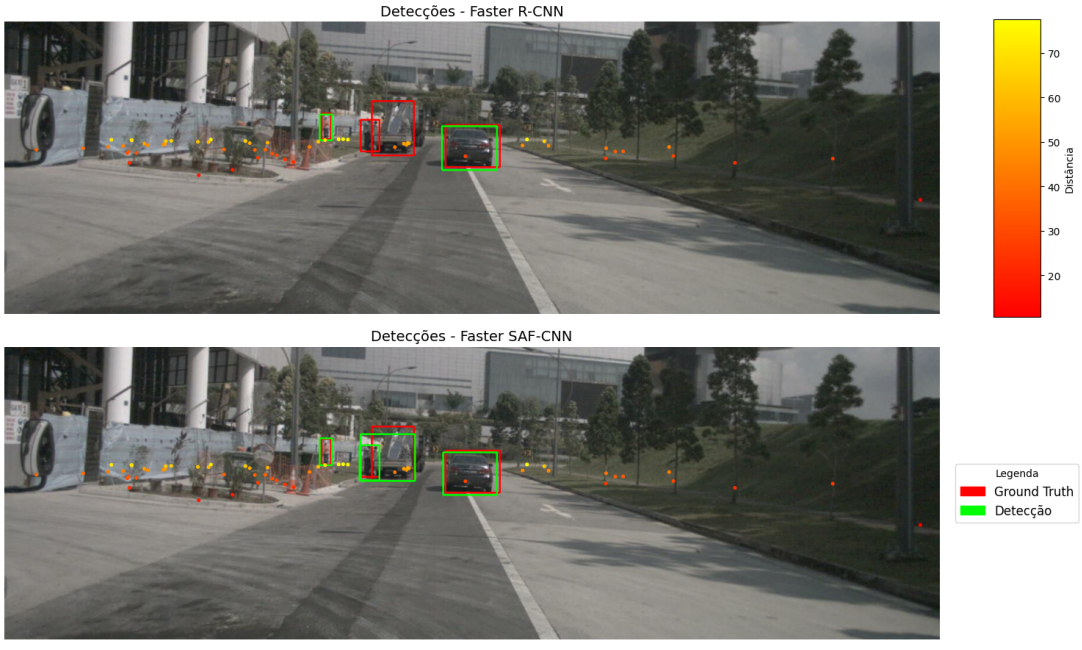

Figura 30 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

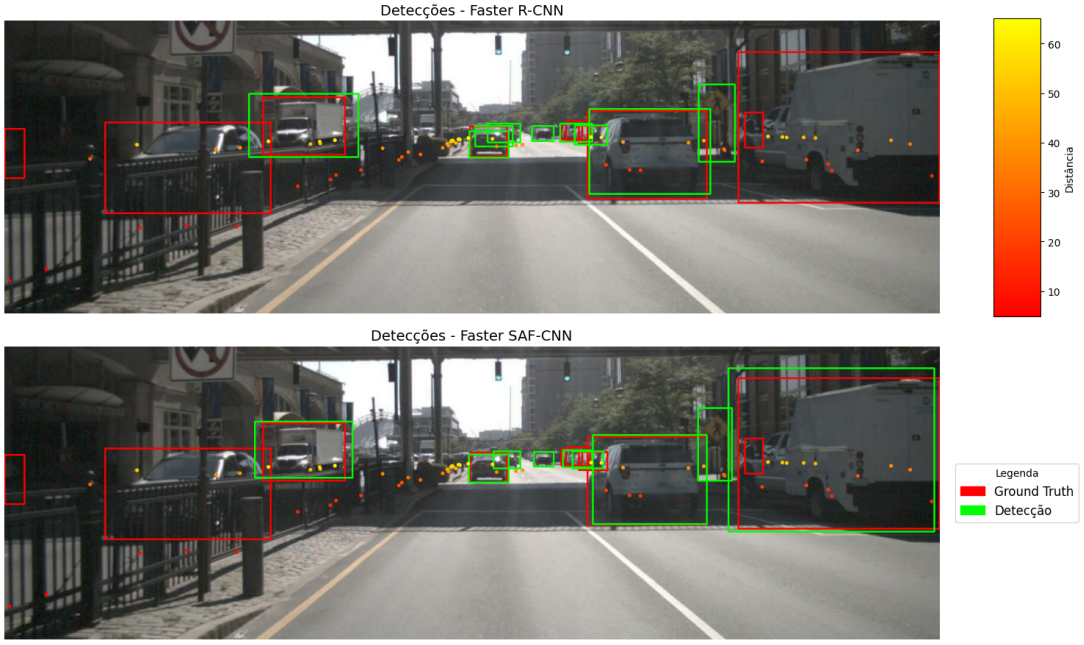

Figura 31 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

5 Conclusões

A percepção precisa do ambiente é um requisito fundamental para a navegação segura de veículos autônomos e motivou a realização deste trabalho, cujo objetivo foi explorar estratégias de fusão sensorial no contexto automotivo, com o uso de algoritmos de detecção baseados em redes neurais convolucionais. A fusão entre sensores de câmera e radar apresenta vantagens significativas, uma vez que esses dispositivos são complementares e amplamente utilizados na indústria automotiva. Essa combinação permite aliar a alta resolução lateral proporcionada pelas câmeras à robustez do radar, especialmente em condições climáticas adversas ou de baixa luminosidade. Além disso, o radar possui menor custo de produção em comparação ao sensor LiDAR, tornando-se uma alternativa atrativa para aplicações em larga escala.

A fundamentação teórica teve início com a apresentação dos Sistemas Avançados de Assistência ao Condutor (ADAS), que contribuem para a segurança veicular ao oferecerem suporte progressivo à direção. Esses sistemas se relacionam diretamente à classificação da norma SAE J3016, que define seis níveis de automação, do controle totalmente manual à condução autônoma plena. Com o avanço desses níveis, cresce a necessidade de mecanismos de percepção ambiental precisos, razão pela qual foram analisados os principais sensores empregados em veículos autônomos, como câmeras, radares e LiDARs, abordando suas características, tipos e funcionalidades. Foram também descritas as três categorias clássicas de fusão sensorial presentes na literatura, com ênfase nas estratégias que combinam dados de câmeras e radares. Por fim, a última seção tratou dos detectores baseados em redes neurais convolucionais, destacando os algoritmos tradicionais das arquiteturas de um e dois estágios.

Para identificar avanços e lacunas, foi realizada uma análise temática da literatura, estruturada em torno de quatro perguntas de pesquisa, distribuídas em três temas principais e um tópico complementar. O primeiro tema tratou da fusão sensorial e sua aplicação voltada para sistemas automotivos. A principal descoberta foi o papel essencial dessa tecnologia na integração de dados complementares e redundantes, o que resulta em maior precisão, robustez e confiabilidade dos sistemas de percepção. A fusão sensorial mostrou-se indispensável para garantir a segurança e a eficiência exigidas pela condução autônoma.

O segundo tema explorou os benefícios e limitações dos sensores câmera, radar e LiDAR para detecção de objetos. Verificou-se que a combinação entre câmeras e radares representa uma solução eficaz e economicamente viável. As câmeras destacam-se na interpretação visual e na detecção de pedestres, enquanto os radares oferecem desempenho superior em condições adversas e fornecem informações cruciais de profundidade e velocidade, evidenciando a complementaridade entre ambos.

Por fim, o terceiro tema tratou dos requisitos para aplicar técnicas modernas de visão computacional à fusão sensorial. Verificou-se que o uso de redes neurais multissensoriais, aliado à transferência de aprendizado, contribui significativamente para a robustez dos modelos que utilizam apenas câmeras. Contudo, a construção de bases de dados multissensoriais com anotações precisas e calibração entre sensores representa um desafio considerável. Nesse sentido, a disponibilidade de conjuntos de dados públicos desempenha papel crucial no avanço da área, permitindo testes comparáveis e a validação de novas metodologias.

Para responder à quarta pergunta, foi incluída uma seção específica dedicada às técnicas de fusão entre câmera e radar. Nessa etapa, foram discutidas as principais arquiteturas propostas na literatura, analisando-as com base nas métricas mais relevantes para a tarefa de detecção 2D de objetos, como Average Precision (AP) e Average Recall (AR). Essa análise permitiu identificar padrões de projeto, destacar as abordagens mais robustas e fornecer uma base sólida para futuras implementações e aprimoramentos.

Na etapa experimental utilizou a base de dados nuScenes, convertida para o formato COCO, para avaliar duas abordagens distintas de fusão sensorial. A primeira abordagem consistiu na fusão em nível de dados, utilizando a arquitetura Radar Region Proposal Network (RRPN). A segunda abordou a fusão em nível de características, por meio do módulo Spatial Attention Fusion (SAF).

A fusão em nível de dados (RRPN), na qual o radar atua como sensor principal, apresentou desempenho inferior ao modelo baseado exclusivamente em câmera. Essa limitação está relacionada ao fato de que muitos objetos anotados na base de dados, como bicicletas, motocicletas e pedestres, frequentemente não possuem pontos de radar associados. Nossa análise indica que mais da metade das bicicletas e motocicletas não apresentam pontos de radar, enquanto apenas 20% dos pedestres têm ao menos um, comprometendo a detecção, pois a ausência de retorno do radar impede a geração de regiões de interesse para a câmera

Em contrapartida, a abordagem de fusão em nível de características baseada no módulo SAF mostrou-se significativamente mais eficaz. Os resultados indicaram ganhos nas métricas de: 1,64% em AP75, 0,96% em AP50, 0,80% em AR e 1,36% em APs, destacando a melhoria na localização precisa das caixas delimitadoras e na detecção de pequenos objetos. A análise do tempo de execução revelou que a maior parte do processamento adicional ocorreu na etapa de geração de propostas, sugerindo que otimizações nessa etapa são necessárias para reduzir o tempo total do modelo. Apesar disso, os avanços obtidos ressaltam o potencial da integração entre as informações visuais da câmera e as medições do radar como estratégia eficaz para aprimorar a percepção em sistemas ADAS, ampliando a capacidade da rede de extrair contextos relevantes mesmo em cenários com oclusões parciais ou neblina. Para atingir tempos próximos ao processamento em tempo real, detectores rápidos e configuráveis, como YOLO, são mais indicados, permitindo ajustar o *trade-off* entre precisão e velocidade.

Nesse contexto, os dados de radar, embora esparsos e sujeitos a ruídos decorrentes de objetos irrelevantes e reflexões no solo, mostraram-se eficazes na detecção de objetos mesmo quando se utiliza apenas a projeção de suas detecções no plano da imagem. Considerando as capacidades ainda subexploradas desse sensor, propõe-se uma investigação mais abrangente, incorporando informações adicionais, como a utilização de atributos de velocidade e direção de movimento dos objetos para rastreamento multiobjeto, o que permite maior robustez frente a oclusões temporárias. Paralelamente, planeja-se abordar o elevado tempo de processamento observado no modelo SAF, especialmente na etapa de geração de propostas, e explorar otimizações no gerador de imagens de radar, atualmente implementado em Python. Também será investigada a influência dos valores elevados nos mapas de características, buscando o equilíbrio entre precisão e velocidade. Nesse contexto, detectores rápidos e configuráveis, como YOLO, poderão ser explorados para atingir tempos próximos ao processamento em tempo real, mantendo a eficácia da fusão multissensorial e promovendo sistemas mais robustos e confiáveis para percepção ambiental multissensorial.

Apesar dos avanços obtidos, o trabalho enfrentou limitações importantes relacionadas ao ambiente de desenvolvimento e aos dados utilizados. O Detectron2 foi adotado inicialmente devido à sua modularidade, suporte a múltiplos detectores e ampla documentação, o que facilitou o ajuste do primeiro modelo de fusão. No entanto, essa escolha impôs restrições técnicas, pois o framework não era compatível com versões mais recentes do CUDA e do PyTorch, inviabilizando o uso dos computadores do laboratório do campus, equipados com hardware mais moderno. Como resultado, o desenvolvimento precisou continuar em um ambiente pessoal, o que limitou a paralelização do treinamento em múltiplas GPUs, reduziu os recursos computacionais disponíveis e aumentou significativamente o tempo de treinamento dos modelos. Do ponto de vista dos dados, a base nuScenes, que é amplamente utilizada na literatura e composta por sensores variados, observou-se uma distribuição limitada de amostras em condições adversas, com apenas 4,7% dos objetos presentes em cenas noturnas e 18,6% em condições de chuva, dificultando a avaliação dos modelos nessas condições. Além disso, as anotações são realizadas apenas quando há pelo menos um ponto de radar ou LiDAR associado ao objeto, o que leva à ausência de rótulos em muitos casos visualmente evidentes, fazendo com que diversas detecções corretas sejam penalizadas por não encontrarem correspondência nas anotações.

Referências

ABDU, F. J.; ZHANG, Y.; FU, M.; LI, Y.; DENG, Z. **Application of deep learning on millimeter-wave radar signals: A review**. v. 21. MDPI AG, mar. 2021. P. 1–46. DOI: 10.3390/s21061951. Citado nas pp. 22, 40.

AG, C. **Radar Sensor for Intelligent Solutions in Industrial Applications**. 2017. https://www.continental.com/en/press/press-releases/radar-sensor-for-intelligent-solutions-in-industrial-applications/. Acesso em: 24 set. 2025. Disponível em: <https://www.continental.com/en/press/press-releases/radar-sensor-for-intelligent-solutions-in-industrial-applications/>. Citado na p. 22.

ALTENDORFER, R.; WIRKERT, S.; HEINRICHS-BARTSCHER, S. Sensor Fusion as an Enabling Technology for Safety-critical Driver Assistance Systems. **SAE International Journal of Passenger Cars - Electronic and Electrical Systems**, v. 3, p. 2010-01–2339, 2 out. 2010. ISSN 1946-4622. DOI: 10.4271/2010-01-2339. Citado na p. 36.

BARNES, D.; GADD, M.; MURCUTT, P.; NEWMAN, P.; POSNER, I. The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset. **arXiv preprint arXiv: 1909.01300**, 2019. Disponível em: <https://arxiv.org/pdf/1909.01300>. Citado na p. 42.

CAESAR, H.; BANKITI, V.; LANG, A. H.; VORA, S.; LIONG, V. E.; XU, Q.; KRISHNAN, A.; PAN, Y.; BALDAN, G.; BEIJBOM, O. nuScenes: A Multimodal Dataset for Autonomous Driving. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, jun. 2020. DOI: 10.1109/cvpr42600.2020.01164. Disponível em: <http://dx.doi.org/10.1109/CVPR42600.2020.01164>. Citado nas pp. 42, 46, 47, 49, 52.

CHANG, S.; ZHANG, Y.; ZHANG, F.; ZHAO, X.; HUANG, S.; FENG, Z.; WEI, Z. Spatial Attention Fusion for Obstacle Detection Using MmWave Radar and Vision Sensor. **Sensors**, MDPI AG, v. 20, n. 4, p. 956, fev. 2020. DOI: 10.3390/s20040956. Disponível em: <https://doi.org/10.3390/s20040956>. Citado nas pp. 24, 25, 43, 44, 46, 57, 58.

CHAVEZ-GARCIA, R. O.; BURLET, J.; VU, T.-D.; AYCARD, O. Frontal object perception using radar and mono-vision. In: 2012 IEEE Intelligent Vehicles Symposium. IEEE, jun. 2012. DOI: 10.1109/ivs.2012.6232307. Disponível em: <https://doi.org/10.1109/ivs.2012.6232307>. Citado na p. 38.

DARMS, M.; FOELSTER, F.; SCHMIDT, J.; FROEHLICH, D.; ECKERT, A. Data Fusion Strategies in Advanced Driver Assistance Systems. **SAE International Journal of Passenger Cars - Electronic and Electrical Systems**, v. 3, p. 2010-01–2337, 2 out. 2010. ISSN 1946-4622. DOI: 10.4271/2010-01-2337. Citado nas pp. 15, 36.

DIMITRIEVSKI, M.; JACOBS, L.; VEELAERT, P.; PHILIPS, W. People Tracking by Cooperative Fusion of RADAR and Camera Sensors. In: p. 509–514. ISBN 978-1-5386-7024-8. DOI: 10.1109/ITSC.2019.8917238. Citado nas pp. 20, 39.

FENG, D.; HAASE-SCHUTZ, C.; ROSENBAUM, L.; HERTLEIN, H.; GLASER, C.; TIMM, F.; WIESBECK, W.; DIETMAYER, K. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. **IEEE Transactions on Intelligent Transportation Systems**, Institute of Electrical e Electronics Engineers (IEEE), v. 22, n. 3, p. 1341–1360, mar. 2021. ISSN 1558-0016. DOI: 10.1109/tits.2020.2972974. Disponível em: <http://dx.doi.org/10.1109/tits.2020.2972974>. Citado nas pp. 28, 41.

GIRSHICK, R. Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, dez. 2015. DOI: 10.1109/iccv.2015.169. Disponível em: <http://dx.doi.org/10.1109/ICCV.2015.169>. Citado nas pp. 26, 27, 57.

GIRSHICK, R.; DONAHUE, J.; DARRELL, T.; MALIK, J. **Rich feature hierarchies for accurate object detection and semantic segmentation**. arXiv, 2013. DOI: 10.48550/ARXIV.1311.2524. Disponível em: <https://arxiv.org/abs/1311.2524>. Citado na p. 25.

HAN, S.; WANG, X.; XU, L.; SUN, H.; ZHENG, N. Frontal object perception for Intelligent Vehicles based on radar and camera fusion. In: 2016-August, p. 4003–4008. ISBN 9789881563910. DOI: 10.1109/ChiCC.2016.7553978. Citado na p. 44.

HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R. **Mask R-CNN**. arXiv, 2017. DOI: 10.48550/ARXIV.1703.06870. Disponível em: <https://arxiv.org/abs/1703.06870>. Citado na p. 26.

HONDA, H. **Digging into Detectron 2 — Part 2: Feature Pyramid Network**. Jan. 2020. Acessado em: Mar. 9, 2025. Disponível em: <https://medium.com/@hirotoschwert/digging-into-detectron-2-part-2-dd6e8b0526e>. Citado na p. 61.

JAHROMI, B. S.; TULABANDHULA, T.; CETIN, S. Real-Time Hybrid Multi-Sensor Fusion Framework for Perception in Autonomous Vehicles. **Sensors**, MDPI AG, v. 19, n. 20, p. 4357, out. 2019. DOI: 10.3390/s19204357. Disponível em: <https://doi.org/10.3390/s19204357>. Citado na p. 22.

JIANG, Q.; ZHANG, L.; MENG, D. Target Detection Algorithm Based on MMW Radar and Camera Fusion. In: p. 1–6. ISBN 978-1-5386-7024-8. DOI: 10.1109/ITSC.2019.8917504. Citado nas pp. 38, 44.

JIAO, L.; ZHANG, F.; LIU, F.; YANG, S.; LI, L.; FENG, Z.; QU, R. A Survey of Deep Learning-Based Object Detection. **IEEE Access**, Institute of Electrical e Electronics Engineers (IEEE), v. 7, p. 128837–128868, 2019. ISSN 2169-3536. DOI: 10.1109/access.2019.2939201. Disponível em: <http://dx.doi.org/10.1109/ACCESS.2019.2939201>. Citado nas pp. 25, 27, 29, 30, 32.

JOHN, A.; MEVA, D. D. A Comparative Study of Various Object Detection Algorithms and Performance Analysis. **INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING**, v. 8, p. 158–163, out. 2020. DOI: 10.26438/ijcse/v8i10.158163. Citado na p. 27.

JOHN, V.; NITHILAN, M. K.; MITA, S.; TEHRANI, H.; SUDHEESH, R. S.; LALU, P. P. SO-Net: Joint Semantic Segmentation and Obstacle Detection Using Deep Fusion of Monocular Camera and Radar. In: 11994 LNCS, p. 138–148. ISBN 9783030397692. DOI: 10.1007/978-3-030-39770-8_11. Citado nas pp. 43, 44.

JOHN, V.; MITA, S. RVNet: Deep Sensor Fusion of Monocular Camera and Radar for Image-Based Obstacle Detection in Challenging Environments. In: LECTURE Notes in Computer Science. Springer International Publishing, 2019. P. 351–364. ISBN 9783030348793. DOI: 10.1007/978-3-030-34879-3_27. Disponível em: <http://dx.doi.org/10.1007/978-3-030-34879-3_27>. Citado nas pp. 43, 44.

KANG, D.; KUM, D. Camera and Radar Sensor Fusion for Robust Vehicle Localization via Vehicle Part Localization. **IEEE Access**, Institute of Electrical e Electronics Engineers Inc., v. 8, p. 75223–75236, 2020. ISSN 21693536. DOI: 10.1109/ACCESS.2020.2985075. Citado nas pp. 39, 44.

KHANAM, R.; HUSSAIN, M. **YOLOv11: An Overview of the Key Architectural Enhancements**. arXiv, 2024. DOI: 10.48550/ARXIV.2410.17725. Disponível em: <https://arxiv.org/abs/2410.17725>. Citado na p. 31.

KHANAM, R.; HUSSAIN, M.; HILL, R.; ALLEN, P. A Comprehensive Review of Convolutional Neural Networks for Defect Detection in Industrial Applications. **IEEE Access**, v. 12, p. 94250–94295, 2024. DOI: 10.1109/ACCESS.2024.3425166. Citado nas pp. 25, 29, 31.

KHARAZI, D. **YOLO Algorithm**. 2025. https://dkharazi.github.io/notes/ml/cnn/yolo. Acessado em: 01 de janeiro de 2025. Citado na p. 30.

KIM, G.; PARK, Y. S.; CHO, Y.; JEONG, J.; KIM, A. MulRan: Multimodal Range Dataset for Urban Place Recognition. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). 2020. P. 6246–6253. DOI: 10.1109/ICRA40945.2020.9197298. Citado na p. 42.

KIM, Y.; KIM, S.; CHOI, J. W.; KUM, D. **CRAFT: Camera-Radar 3D Object Detection with Spatio-Contextual Fusion Transformer**. arXiv, 2022. DOI: 10.48550/ARXIV.2209.06535. Disponível em: <https://arxiv.org/abs/2209.06535>. Citado na p. 38.

KOCIC, J.; JOVICIC, N.; DRNDAREVIC, V. Sensors and Sensor Fusion in Autonomous Vehicles. In: 2018 26th Telecommunications Forum (TELFOR). IEEE, nov. 2018. DOI: 10.1109/telfor.2018.8612054. Disponível em: <https://doi.org/10.1109/telfor.2018.8612054>. Citado na p. 15.

KUMAR, R.; JAYASHANKAR, S. Radar and Camera Sensor Fusion with ROS for Autonomous Driving. In: p. 568–573. ISBN 978-1-7281-0899-5. DOI: 10.1109/ICIIP47207.2019.8985782. Citado nas pp. 22, 37.

LI, L. Q.; XIE, Y. L. A Feature Pyramid Fusion Detection Algorithm Based on Radar and Camera Sensor. In: 2020-December, p. 366–370. ISBN 9781728144795. DOI: 10.1109/ICSP48669.2020.9320985. Citado nas pp. 38, 39, 43, 44.

LI, Z.; DONG, Y.; SHEN, L.; LIU, Y.; PEI, Y.; YANG, H.; ZHENG, L.; MA, J. Development and challenges of object detection: A survey. **Neurocomputing**, Elsevier BV, v. 598, p. 128102, set. 2024. ISSN 0925-2312. DOI: 10.1016/j.neucom.2024.128102. Disponível em: <http://dx.doi.org/10.1016/j.neucom.2024.128102>. Citado na p. 26.

LIN, T.-Y.; DOLLÁR, P.; GIRSHICK, R.; HE, K.; HARIHARAN, B.; BELONGIE, S. **Feature Pyramid Networks for Object Detection**. arXiv, 2016. DOI: 10.48550/ARXIV.1612.03144. Disponível em: <https://arxiv.org/abs/1612.03144>. Citado na p. 32.

LIN, T.-Y.; GOYAL, P.; GIRSHICK, R.; HE, K.; DOLLAR, P. Focal Loss for Dense Object Detection. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Institute of Electrical e Electronics Engineers (IEEE), v. 42, n. 2, p. 318–327, fev. 2020. ISSN 1939-3539. DOI: 10.1109/tpami.2018.2858826. Disponível em: <http://dx.doi.org/10.1109/TPAMI.2018.2858826>. Citado nas pp. 26, 31, 32.

LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; BOURDEV, L.; GIRSHICK, R.; HAYS, J.; PERONA, P.; RAMANAN, D.; ZITNICK, C. L.; DOLLÁR, P. **Microsoft COCO: Common Objects in Context**. arXiv, 2014. DOI: 10.48550/ARXIV.1405.0312. Disponível em: <https://arxiv.org/abs/1405.0312>. Citado na p. 49.

LIU, Y.; CHANG, S.; WEI, Z.; ZHANG, K.; FENG, Z. Fusing mmWave Radar With Camera for 3-D Detection in Autonomous Driving. **IEEE Internet of Things Journal**, Institute of Electrical e Electronics Engineers Inc., v. 9, p. 20408–20421, 20 out. 2022. ISSN 23274662. DOI: 10.1109/JIOT.2022.3175375. Citado nas pp. 16, 21, 39.

LIU, Z.; CAI, Y.; WANG, H.; CHEN, L.; GAO, H.; JIA, Y.; LI, Y. Robust Target Recognition and Tracking of Self-Driving Cars With Radar and Camera Information Fusion Under Severe Weather Conditions. **IEEE Transactions on Intelligent Transportation Systems**, Institute of Electrical e Electronics Engineers Inc., v. 23, p. 6640–6653, 7 jul. 2022. ISSN 15580016. DOI: 10.1109/TITS.2021.3059674. Citado nas pp. 15, 16, 23, 37, 44.

LU, Y.; XUE, Z.; XIA, G.-S.; ZHANG, L. A survey on vision-based UAV navigation. **Geospatial Information Science**, Informa UK Limited, v. 21, n. 1, p. 21–32, jan. 2018. ISSN 1993-5153. DOI: 10.1080/10095020.2017.1420509. Disponível em: <http://dx.doi.org/10.1080/10095020.2017.1420509>. Citado na p. 21.

MEYER, M. Automotive Radar Dataset for Deep Learning Based 3D Object Detection. In. Citado na p. 42.

MICHAELIS, C.; MITZKUS, B.; GEIRHOS, R.; RUSAK, E.; BRINGMANN, O.; ECKER, A. S.; BETHGE, M.; BRENDEL, W. **Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming**. arXiv, 2019. DOI: 10.48550/ARXIV.1907.07484. Disponível em: <https://arxiv.org/abs/1907.07484>. Citado na p. 38.

NABATI, R.; HARRIS, L.; QI, H. CFTrack: Center-based Radar and Camera Fusion for 3D Multi-Object Tracking. In: p. 243–248. ISBN 9781665479219. DOI: 10.1109/IVWorkshops54471.2021.9669223. Citado na p. 16.

NABATI, R.; QI, H. CenterFusion: Center-based radar and camera fusion for 3d object detection. In: p. 1526–1535. ISBN 9780738142661. DOI: 10.1109/WACV48630.2021.00157. Citado na p. 22.

NABATI, R.; QI, H. Radar-Camera Sensor Fusion for Joint Object Detection and Distance Estimation in Autonomous Vehicles. arXiv, 2020. DOI: 10.48550/ARXIV.2009.08428. Citado nas pp. 25, 43, 44.

NABATI, R.; QI, H. RRPN: Radar Region Proposal Network for Object Detection in Autonomous Vehicles. arXiv, 2019. DOI: 10.48550/ARXIV.1905.00526. Disponível em: <https://arxiv.org/abs/1905.00526>. Citado nas pp. 43, 44, 50, 57.

NGUYEN, N.-D.; DO, T.; NGO, T. D.; LE, D.-D. An Evaluation of Deep Learning Methods for Small Object Detection. **Journal of Electrical and Computer Engineering**, Hindawi Limited, v. 2020, p. 1–18, abr. 2020. ISSN 2090-0155. DOI: 10.1155/2020/3189691. Disponível em: <http://dx.doi.org/10.1155/2020/3189691>. Citado na p. 26.

NOBIS, F.; GEISSLINGER, M.; WEBER, M.; BETZ, J.; LIENKAMP, M. **A Deep Learning-based Radar and Camera Sensor Fusion Architecture for Object Detection**. arXiv, 2020. DOI: 10.48550/ARXIV.2005.07431. Disponível em: <https://arxiv.org/abs/2005.07431>. Citado nas pp. 43, 44.

OTTO, C.; GERBER, W.; LEON, F. P.; WIRNITZER, J. A Joint Integrated Probabilistic Data Association Filter for pedestrian tracking across blind regions using monocular camera and radar. In: 2012 IEEE Intelligent Vehicles Symposium. IEEE, jun. 2012. DOI: 10.1109/ivs.2012.6232228. Disponível em: <http://dx.doi.org/10.1109/IVS.2012.6232228>. Citado na p. 18.

PARK, J.; YU, W. A sensor fused rear cross traffic detection system using transfer learning. **Sensors**, MDPI, v. 21, 18 set. 2021. ISSN 14248220. DOI: 10.3390/s21186055. Citado nas pp. 18, 40, 44.

REDMON, J.; DIVVALA, S.; GIRSHICK, R.; FARHADI, A. You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, jun. 2016. P. 779–788. DOI: 10.1109/cvpr.2016.91. Disponível em: <http://dx.doi.org/10.1109/CVPR.2016.91>. Citado nas pp. 26, 29.

REN, S.; HE, K.; GIRSHICK, R.; SUN, J. **Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks**. arXiv, 2015. DOI: 10.48550/ARXIV.1506.01497. Disponível em: <https://arxiv.org/abs/1506.01497>. Citado nas pp. 26, 28, 32, 54, 55, 57, 63.

SAE INTERNATIONAL. **SAE Updates J3016 Automated-Driving Graphic**. Jan. 2019. https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic. Acesso em 20 jul. 2025. Citado na p. 20.

SHEENY, M.; DE PELLEGRIN, E.; MUKHERJEE, S.; AHRABIAN, A.; WANG, S.; WALLACE, A. RADIATE: A Radar Dataset for Automotive Perception in Bad Weather. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, mai. 2021. P. 1–7. DOI: 10.1109/icra48506.2021.9562089. Disponível em: <http://dx.doi.org/10.1109/ICRA48506.2021.9562089>. Citado nas pp. 41, 42.

SULTANA, F.; SUFIAN, A.; DUTTA, P. A Review of Object Detection Models Based on Convolutional Neural Network. In: INTELLIGENT Computing: Image Processing Based Applications. Springer Singapore, 2020. P. 1–16. ISBN 9789811542886. DOI: 10.1007/978-981-15-4288-6_1. Disponível em: <http://dx.doi.org/10.1007/978-981-15-4288-6_1>. Citado nas pp. 25, 26, 28, 29.

TERVEN, J.; CÓRDOVA-ESPARZA, D.-M.; ROMERO-GONZÁLEZ, J.-A. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. **Machine Learning and Knowledge Extraction**, MDPI AG, v. 5, n. 4,

p. 1680–1716, nov. 2023. ISSN 2504-4990. DOI: 10.3390/make5040083. Disponível em: <http://dx.doi.org/10.3390/make5040083>. Citado na p. 30.

TIAN, Y.; YE, Q.; DOERMANN, D. **YOLOv12: Attention-Centric Real-Time Object Detectors**. arXiv, 2025. DOI: 10.48550/ARXIV.2502.12524. Disponível em: <https://arxiv.org/abs/2502.12524>. Citado na p. 31.

WANG, A.; CHEN, H.; LIU, L.; CHEN, K.; LIN, Z.; HAN, J.; DING, G. **YOLOv10: Real-Time End-to-End Object Detection**. arXiv, 2024. DOI: 10.48550/ARXIV.2405.14458. Disponível em: <https://arxiv.org/abs/2405.14458>. Citado na p. 31.

WANG, C.-Y.; YEH, I.-H.; LIAO, H.-Y. M. **YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information**. arXiv, 2024. DOI: 10.48550/ARXIV.2402.13616. Disponível em: <https://arxiv.org/abs/2402.13616>. Citado na p. 31.

WANG, D.; WATKINS, C.; XIE, H. MEMS Mirrors for LiDAR: A Review. **Micromachines**, MDPI AG, v. 11, n. 5, p. 456, abr. 2020. ISSN 2072-666X. DOI: 10.3390/mi11050456. Disponível em: <http://dx.doi.org/10.3390/mi11050456>. Citado na p. 23.

WEI, Z.; ZHANG, F.; CHANG, S.; LIU, Y.; WU, H.; FENG, Z. **MmWave Radar and Vision Fusion for Object Detection in Autonomous Driving: A Review**. v. 22. MDPI, abr. 2022. DOI: 10.3390/s22072542. Citado nas pp. 24, 37, 41.

WU, X.; REN, J.; WU, Y.; SHAO, J. Study on Target Tracking Based on Vision and Radar Sensor Fusion. In: 2018-April. DOI: 10.4271/2018-01-0613. Citado na p. 24.

WU, Y.; KIRILLOV, A.; MASSA, F.; LO, W.-Y.; GIRSHICK, R. **Detectron2**. 2019. https://github.com/facebookresearch/detectron2. Citado nas pp. 45, 57.

YADAV, R.; VIERLING, A.; BERNS, K. Radar + RGB Fusion For Robust Object Detection In Autonomous Vehicle. In: 2020 IEEE International Conference on Image Processing (ICIP). 2020. P. 1986–1990. DOI: 10.1109/ICIP40778.2020.9191046. Citado nas pp. 43, 44.

YAN, G.; LIU, Z.; WANG, C.; SHI, C.; WEI, P.; CAI, X.; MA, T.; LIU, Z.; ZHONG, Z.; LIU, Y.; ZHAO, M.; MA, Z.; LI, Y. OpenCalib: A Multi-sensor Calibration Toolbox for Autonomous Driving. **arXiv preprint arXiv:2205.14087**, 2022. Citado na p. 41.

YEONG, D. J.; VELASCO-HERNANDEZ, G.; BARRY, J.; WALSH, J. **Sensor and sensor fusion technology in autonomous vehicles: A review**. v. 21. MDPI AG, mar. 2021. P. 1–37. DOI: 10.3390/s21062140. Citado nas pp. 19, 21, 23, 24, 41.

YU, Z.; BAI, J.; CHEN, S.; HUANG, L.; BI, X. Camera-Radar Data Fusion for Target Detection via Kalman Filter and Bayesian Estimation. In: 2018-August. DOI: 10.4271/2018-01-1608. Citado na p. 15.

ZHANG, X.; ZHOU, M.; QIU, P.; HUANG, Y.; LI, J. Radar and vision fusion for the real-time obstacle detection and identification. **Industrial Robot**, Emerald Group Holdings Ltd., v. 46, p. 391–395, 3 ago. 2019. ISSN 0143991X. DOI: 10.1108/IR-06-2018-0113. Citado nas pp. 41, 44.

ZHOU, J. A Review of LiDAR sensor Technologies for Perception in Automated Driving. **Academic Journal of Science and Technology**, v. 3, p. 255–261, nov. 2022. DOI: 10.54097/ajst.v3i3.2993. Citado na p. 23.

ZHU, Y.; WANG, T.; ZHU, S. Adaptive Multi-Pedestrian Tracking by Multi-Sensor: Track-to-Track Fusion Using Monocular 3D Detection and MMW Radar. **Remote Sensing**, MDPI, v. 14, 8 abr. 2022. ISSN 20724292. DOI: 10.3390/rs14081837. Citado na p. 21.

ZIEBINSKI, A.; CUPEK, R.; ERDOGAN, H.; WAECHTER, S. A survey of ADAS technologies for the future perspective of sensor fusion. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Springer Verlag, 9876 LNCS, p. 135–146, 2016. ISSN 16113349. DOI: 10.1007/978-3-319-45246-3_13. Citado nas pp. 18, 21, 37.

ZOLLHÖFER, M.; STOTKO, P.; GÖRLITZ, A.; THEOBALT, C.; NIESSNER, M.; KLEIN, R.; KOLB, A. State of the Art on 3D Reconstruction with RGB-D Cameras. **Computer Graphics Forum**, Wiley, v. 37, n. 2, p. 625–652, mai. 2018. ISSN 1467-8659. DOI: 10.1111/cgf.13386. Disponível em: <http://dx.doi.org/10.1111/cgf.13386>. Citado nas pp. 21, 22.

Apêndices

APÊNDICE A – Códigos

A.1 Fluxo dos dados do radar no Detectron2

Código A.1 – Formato do dicionário gerado pelo dados COCO

```
{
    "file_name": "caminho/imagem_01.jpg",
    "height": 900,
    "width": 1600,
    "image_id": 01
    "annotations": [
        {
            "iscrowd": 0,
            "bbox": [ 613.17, 381.07, 262.68, 151.08 ],
            "category_id": 5,
            "bbox_mode": 1
        },
        % Mais anotações aqui
    ],
    "pointcloud": [
        [   1298.83, 650.80, 7.71, -5.00, -2.25 ],
        [   1574.58, 670.32, 6.91, -5.25, -2.00 ],
        % Mais pontos aqui
    ]
}
% Mais imagens
```

Anexos

ANEXO A – Propostas de Região

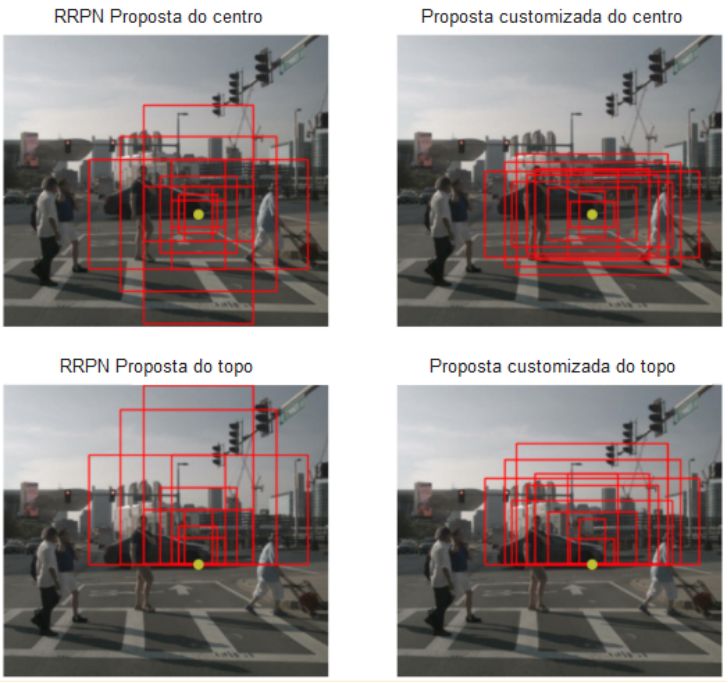

Figura 32 – Âncoras RRPN e customizada centro e topo.

Fonte: Produzido pelo autor com dados de nuScenes

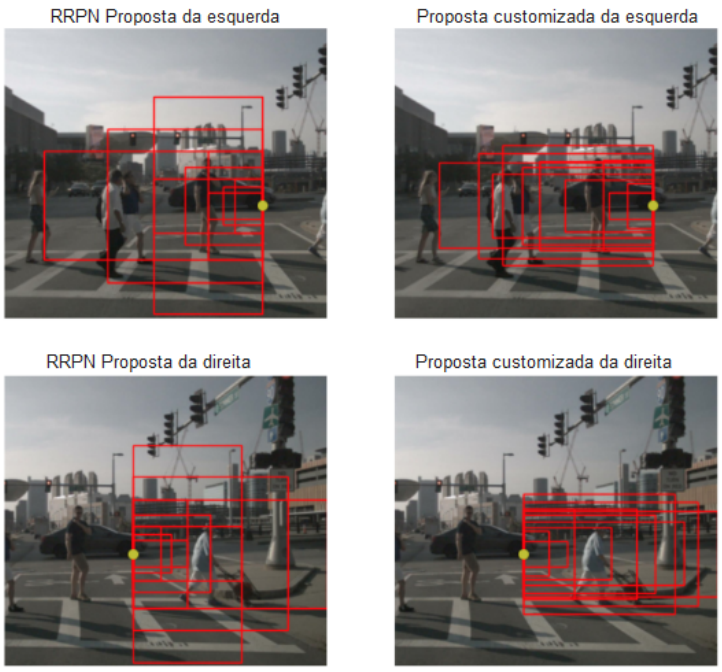

Figura 33 – Âncoras RRPN e customizada esquerda e direita.

Fonte: Produzido pelo autor com dados de nuScenes

ANEXO B – Detecções Faster R-CNN versus modelo com Fusão SAF

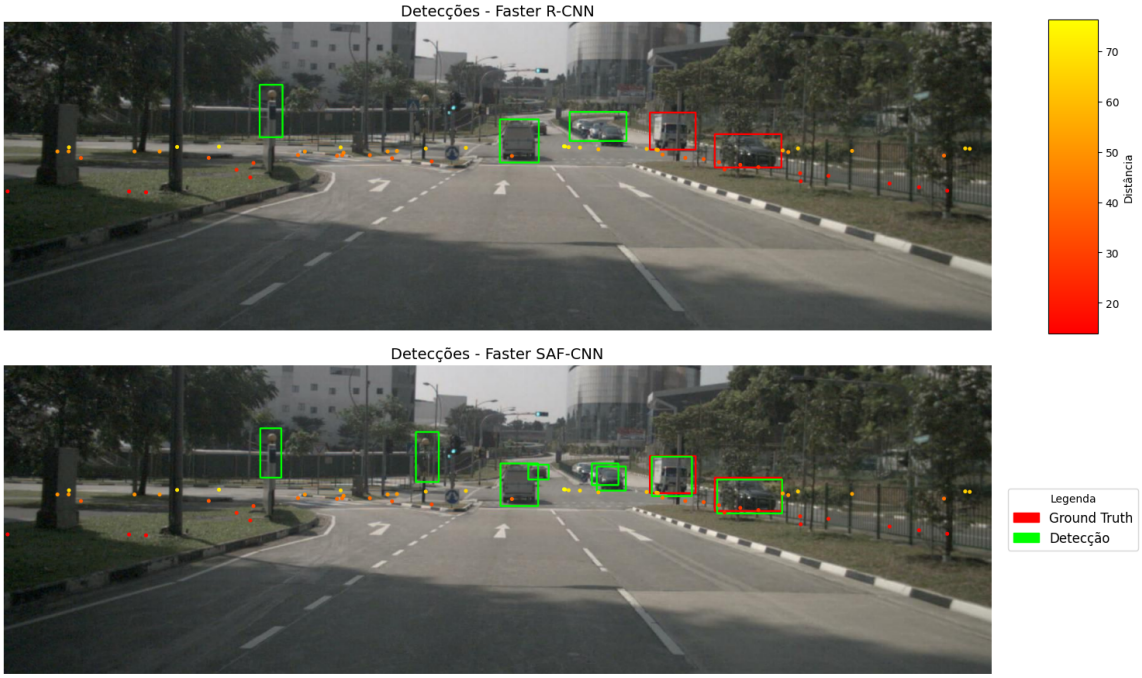

Figura 34 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

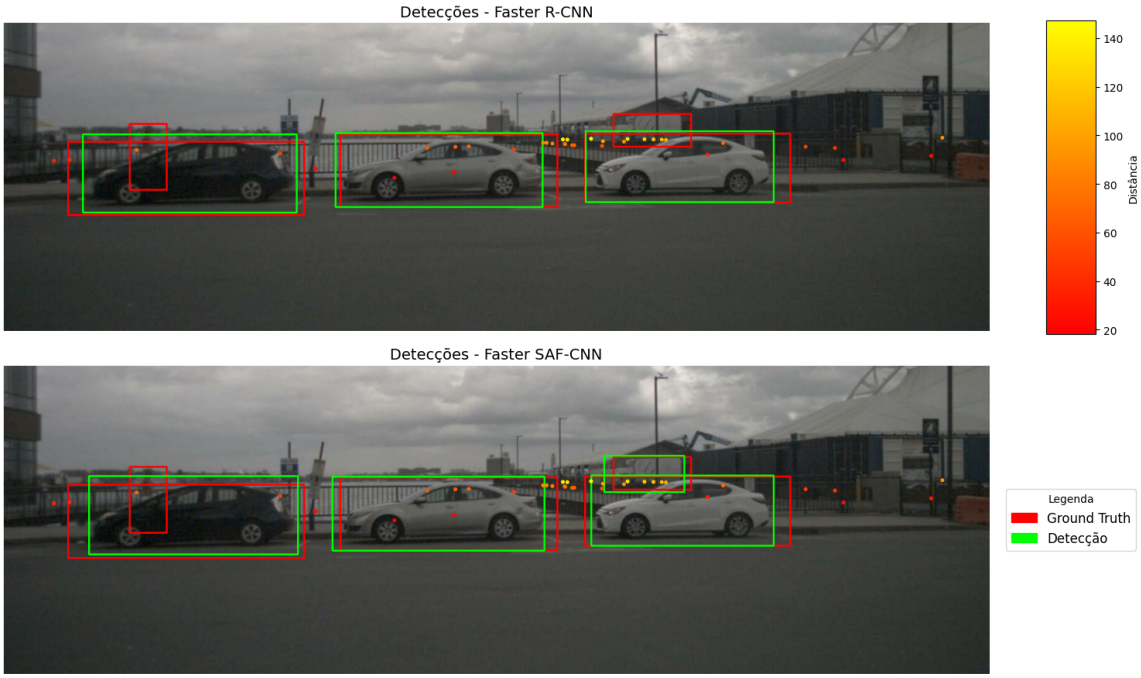

Figura 35 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

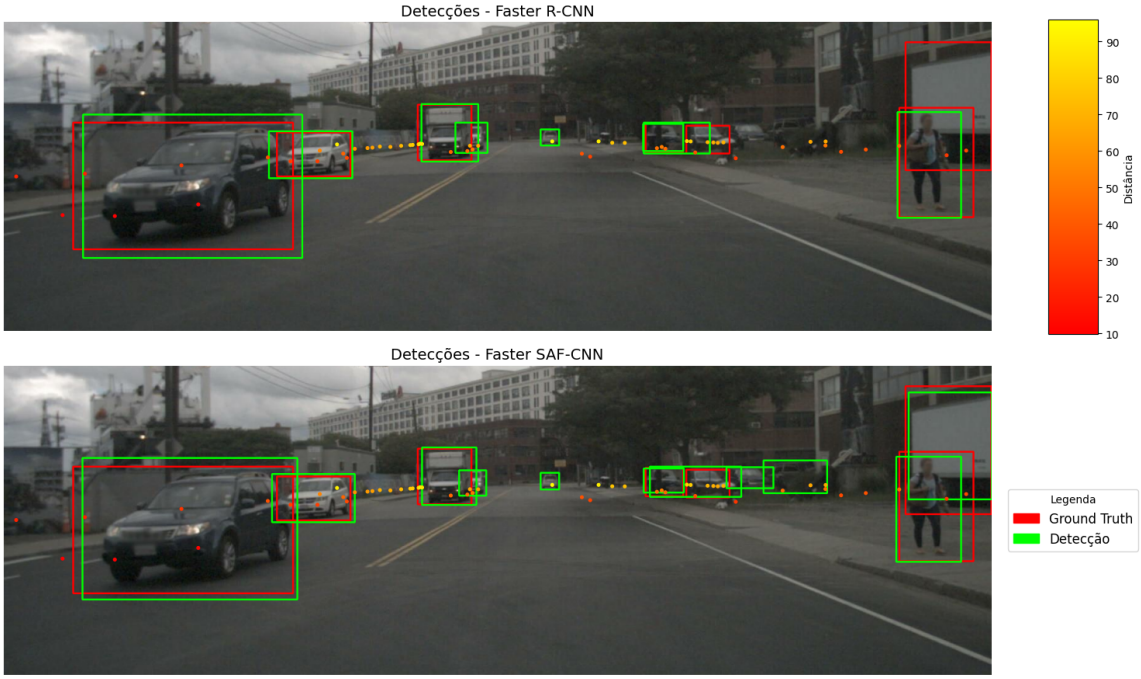

Figura 36 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

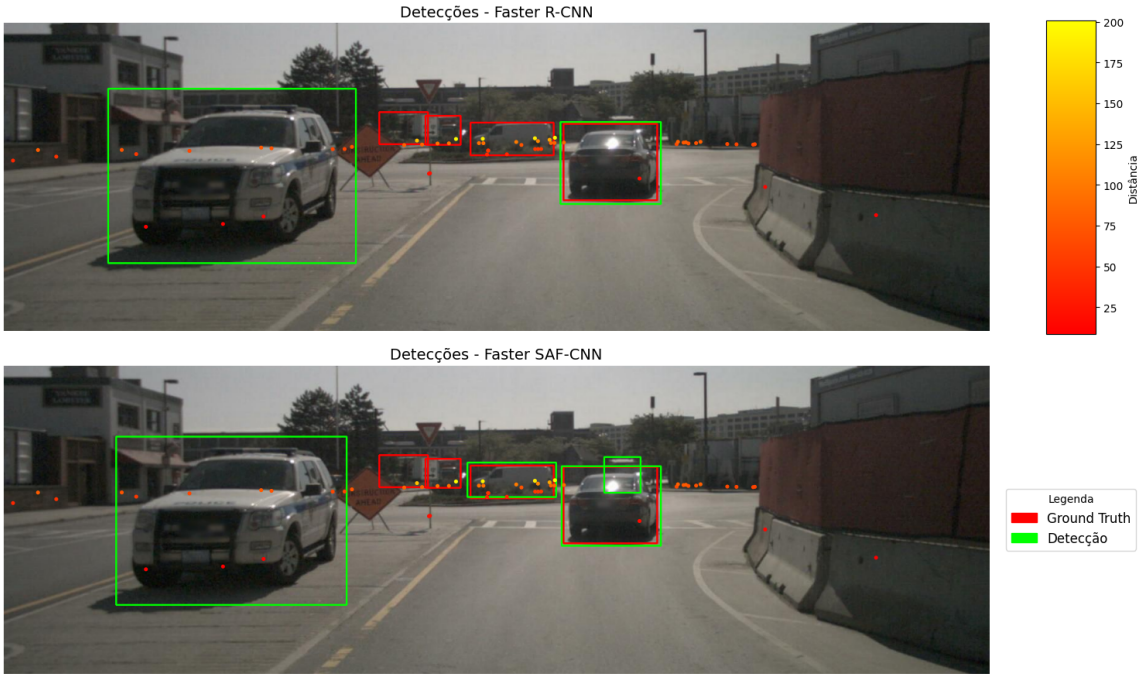

Figura 37 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

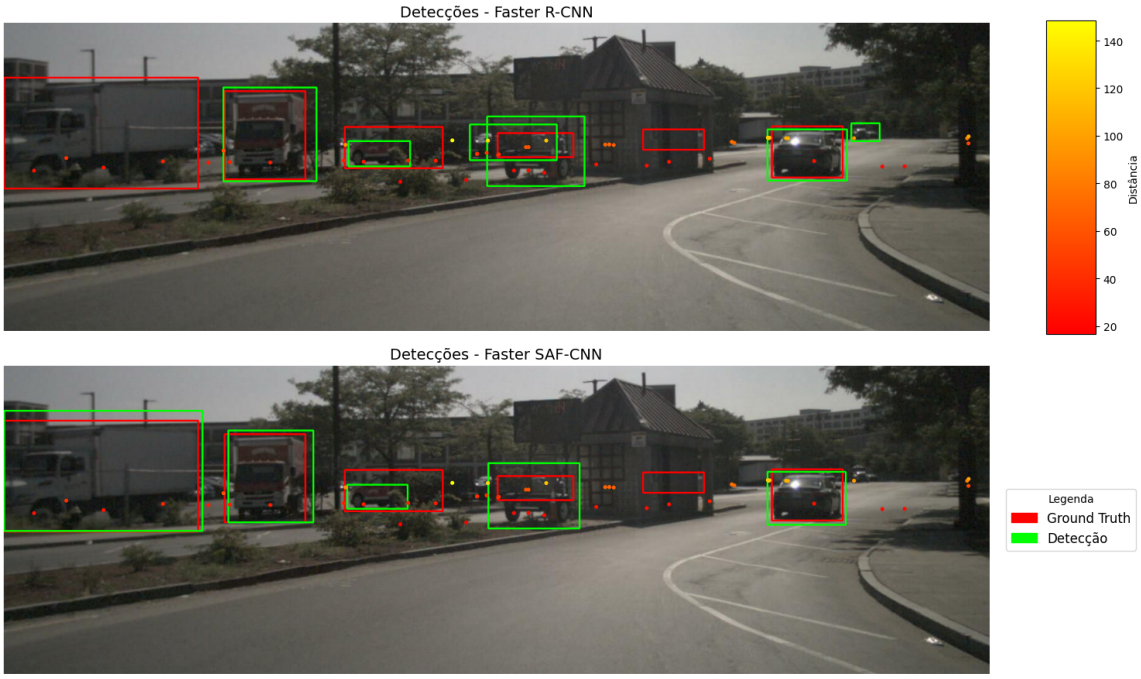

Figura 38 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

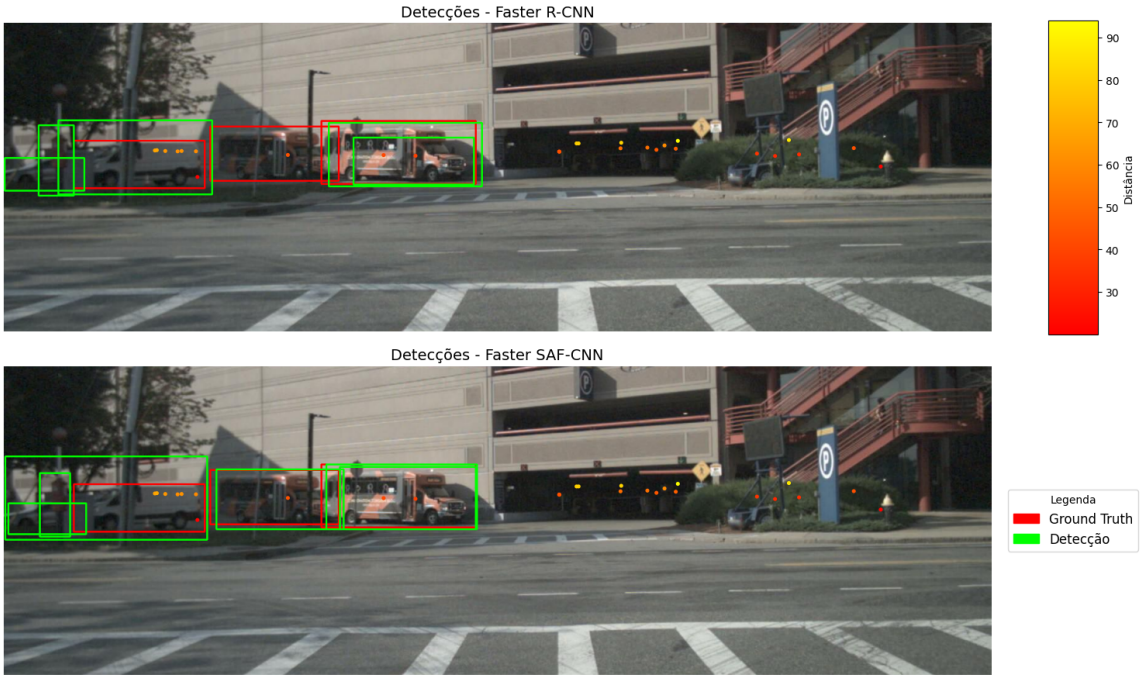

Figura 39 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

Figura 40 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

Figura 41 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes

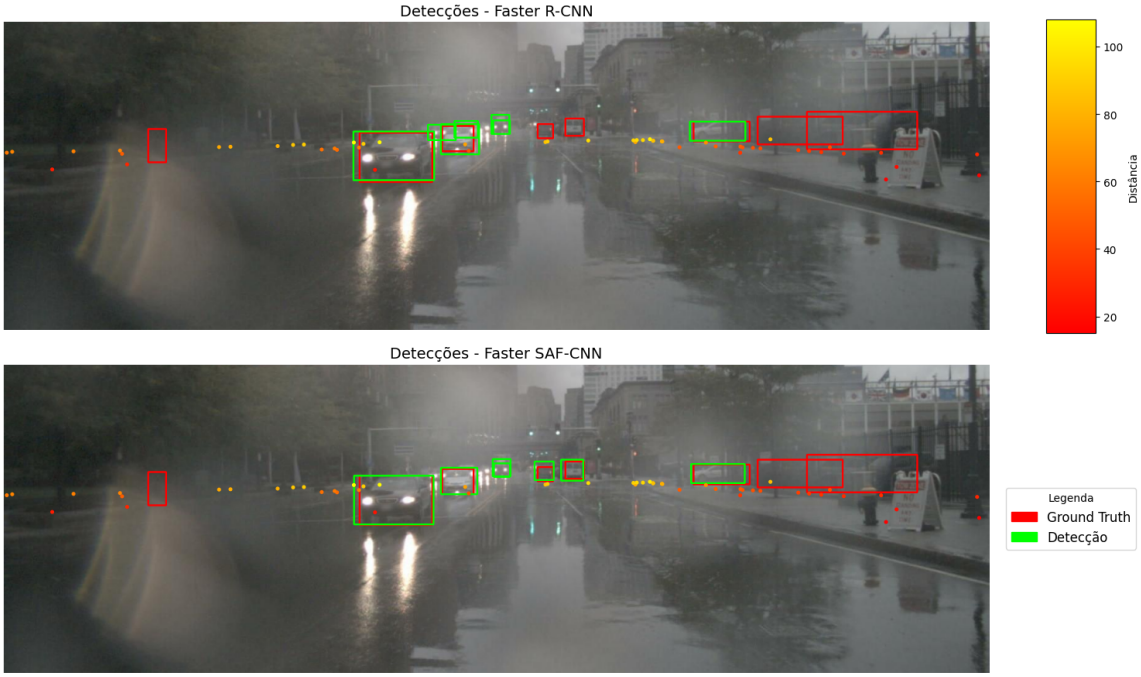

Figura 42 – Comparação da detecção entre os modelos.

Fonte: Produzido pelo autor com dados de nuScenes